Vertrauen in einer digitalen Welt

Eine Betrachtung unter Berücksichtigung von
Neurobiologie und ausgewählten individuellen Unterschieden

Inaugural-Dissertation
zur Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität
zu Bonn

vorgelegt von

Dr.rer.soc.oec.

René Riedl

aus

Steyr, Österreich

Bonn 2025

Gedruckt mit der Genehmigung der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Prof. Dr. Rainer Banse, Institut für Psychologie

(Vorsitzende/Vorsitzender)

Prof. Dr. Martin Reuter, Institut für Psychologie

(Betreuerin/Betreuer und Gutachterin/Gutachter)

Prof. Dr. Henning Gibbons, Institut für Psychologie

(Gutachterin/Gutachter)

PD Dr. Ina Grau, Institut für Psychologie

(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 04.02.2025

Danksagung

Dank gebührt allen voran Univ.-Prof. Dr. Dipl. Psych. Martin Reuter, Leiter der Abteilung für Differentielle & Biologische Psychologie an der Universität Bonn. Im gesamten Dissertationsprozess stand er immer mit Rat und Tat zur Verfügung. Weiter danke ich allen Ko-Autoren. Der wissenschaftliche Austausch und die Zusammenarbeit mit den Ko-Autoren waren nicht nur intellektuell bereichernd, sondern fortwährend auch von Kollegialität und Wertschätzung geprägt. Schließlich danke ich meiner Familie, allen voran meiner Frau Ingrid sowie meinen Söhnen Pascal und Nils, Patrik, Tina, Julian und Fabian, sowie – last but not least – meinen Eltern.

Zusammenfassung

Die Entscheidung, bestimmten Menschen zu vertrauen, anderen jedoch nicht, ist seit jeher wichtig und hat auch evolutionsbiologisch große Bedeutung. Ein wesentliches Merkmal von Vertrauensbeziehungen über fast die gesamte Menschheitsgeschichte war, dass Interaktion von Angesicht zu Angesicht stattfand. Dieses Merkmal hat sich jedoch grundlegend verändert. Durch die zunehmende Verbreitung von Informations- und Kommunikationstechnologien und die damit einhergehende digitale Transformation von Wirtschaft und Gesellschaft haben sich Interaktionsmodi und -verhalten von Menschen dramatisch verändert. Viele Menschen erledigen Aufgaben über das Internet und kommunizieren heutzutage auch vielfach online. Zudem interagieren Menschen zunehmend mit Maschinen, Robotern und virtuellen Agenten. Dieser Trend zur Digitalisierung hat erhebliche Auswirkungen auf die Rolle des Vertrauens in Wirtschaft und Gesellschaft. Analysiert man die Fachliteratur zur Vertrauensforschung, dann fallen insbesondere zwei Aspekte auf: Erstens werden zunehmend mehr die neurobiologischen Grundlagen von Vertrauen erforscht und damit einhergehend hat sich mittlerweile ein umfangreicher Fundus an empirischen Erkenntnissen entwickelt; dieser bezieht sich vor allem auf drei Analyseebenen (Genetik, Hormone, Gehirn), die in Wechselwirkung miteinander stehen. Zweitens können individuelle Unterschiede (z. B. Geschlecht, Persönlichkeit) einen signifikanten Einfluss auf die Vertrauenswahrnehmung und das Vertrauensverhalten haben, wobei diese Unterschiede – zumindest teilweise – auf neurobiologische Unterschiede zurückgeführt werden können. In Anbetracht dieser Situation befasst sich die vorliegende Dissertation mit *Vertrauen in einer digitalen Welt und berücksichtigt dabei die Neurobiologie des Vertrauens sowie ausgewählte individuelle Unterschiede (konkret Geschlecht, Persönlichkeit und die Parkinson-Krankheit).* Die Dissertation setzt sich aus sieben Einzelarbeiten (Publikationen) zusammen; zudem wird ein Grundlagenteil dargeboten, der das Fundament der Einzelarbeiten darstellt. Der wesentliche Beitrag dieser Dissertation liegt in der integrativen Betrachtung der Phänomene „Vertrauen“, „Digitale Welt“ und „Neurobiologie“, wobei zusätzlich der Faktor „Individuelle Unterschiede“ explizite Berücksichtigung findet. Ein solcher Fokus existierte bis zum Zeitpunkt der Publikation der ersten Arbeiten dieser Dissertation nicht – daher wird durch die vorliegende Arbeit ein wirksamer Beitrag geleistet, eine wichtige Forschungslücke zu schließen. Die Verwendung von Digitaltechnologien kann für das Individuum, Organisationen sowie die Gesellschaft insgesamt einen Nutzen stiften (z. B. Informationszugang, Kommunikationsmöglichkeiten, Produktivitätssteigerungen und Innovationspotenzial), birgt jedoch auch Risiken. In welchem Ausmaß und in welcher Weise die Vorteile wirksam werden, hängt maßgeblich vom Vertrauen in die Technologien ab. Vertrauen im technologischen Kontext besser zu verstehen ist somit wichtig und diese Dissertation hat den Anspruch, einen Erkenntnisbeitrag auf diesem Gebiet zu leisten. Schließlich werden die Limitationen der hier vorgestellten Forschungsarbeiten dargelegt und mögliche künftige Forschungsausrichtungen werden transparent gemacht.

Inhaltsverzeichnis

1. Einleitung

Die Entscheidung darüber, welchen Menschen man vertraut, und welchen nicht, ist seit Anbeginn der Menschheit von hoher Bedeutung. Vertrauen in Menschen zu haben, die sich als nicht vertrauenswürdig herausstellen, kann einerseits mit signifikanten Verlusten einhergehen, im schlimmsten Fall sogar mit dem Tod. Andererseits ist Vertrauen eine wesentliche Grundlage für individuelles Wohlbefinden und volkswirtschaftlichen Wohlstand. Ein wesentliches Merkmal von Vertrauensbeziehungen über weite Strecken der Menschheitsgeschichte war, dass Interaktion von Angesicht zu Angesicht stattfand. Dieses Merkmal hat sich jedoch in der jüngsten Phase der Menschheitsgeschichte grundlegend verändert. Durch die zunehmende Verbreitung von Informations- und Kommunikationstechnologien und die damit einhergehende digitale Transformation von Wirtschaft und Gesellschaft haben sich Interaktionsmodi und -verhalten von Menschen dramatisch verändert. Viele Menschen erledigen Aufgaben über das Internet und kommunizieren heutzutage auch vielfach online. Man denke hier an Online Shopping (z. B. Amazon oder eBay) sowie an Kommunikation via Social Media (z. B. Facebook oder Instagram). Zudem interagieren Menschen zunehmend mit Maschinen, Robotern und virtuellen Agenten. Man denke hier an autonom fahrende Autos, Check-in-Terminals auf Flughäfen, Bestellautomaten in Fast-Food-Restaurants, Ticketmaschinen auf Bahnhöfen sowie an jüngste Entwicklungen auf der Basis von Künstlicher Intelligenz (KI) wie ChatGPT sowie ähnliche Technologien wie Sprachassistenten (z. B. Amazon's Alexa oder Apple's Siri).

Dieser Trend zur Digitalisierung hat erhebliche Auswirkungen auf die Rolle des Vertrauens in Wirtschaft und Gesellschaft als auch darauf, wie Vertrauen und Vertrauenswürdigkeit zu konzeptualisieren sind. Unter anderem hängen die Faktoren zur Einschätzung der Vertrauenswürdigkeit eines Interaktionspartners sowie die Prozesse der Vertrauensbildung davon ab, ob Interaktion von Angesicht-zu-Angesicht oder über ein elektronisches Medium (z. B. Social Media) stattfindet, sowie davon, ob ein Interaktionspartner als Mensch oder als Maschine (z. B. Sprachassistent) wahrgenommen wird.

Die enorme Relevanz dieser Veränderungen wird deutlich, wenn man sich (i) den mittlerweile in weiten Teilen der Welt bestehenden sehr hohen Verwendungsgrad von Digitaltechnologien vor Augen führt und (ii) berücksichtigt, dass Vertrauen mit einer Vielzahl positiver Wirkungen einhergeht, was bislang viel häufiger in Situationen klassischer sozialer Interaktion als bei digitaler Interaktion untersucht wurde.

Zum hohen Verwendungsgrad von Digitaltechnologien können unter anderem folgende Zahlen dargeboten werden: Veröffentlichungen von Statista zeigen, dass im Jahr 2022 die Anzahl der weltweiten Internetnutzer die 5-Milliardengrenze durchbrochen hat, das sind rund zwei Drittel der Weltbevölkerung (Statista 2022). Eine Mehrheit dieser Internetnutzer verwendet Social Media (4,76 Milliarden), wobei die konkreten Zahlen für weithin bekannte Plattformen wie folgt angegeben werden (in Milliarden): Facebook (2,96), Youtube (2,51), WhatsApp (2,00), Instagram (2,00), WeChat (1,31), TikTok (1,05), Telegram (0,70), Snapchat (0,64) und Twitter (mittlerweile X) (0,56) (Statista 2023a). Weiter wird darüber berichtet, dass aktuell rund 3,9 Milliarden Menschen ein Smartphone verwenden und allein im Jahr 2022 weltweit 1,2 Milliarden Smartphones verkauft wurden (Statista 2023b). Schließlich zeigen Befunde der International Telecommunication Union (ITU) aus dem Jahr 2022, dass 95 % der Weltbevölkerung Zugang zu mobilem Breitband (3G oder darüber) haben, um darüber ins Internet zu gelangen (ITU 2022).

Zu den positiven Wirkungen von Vertrauen kann angeführt werden, dass die ökonomische Forschung zeigt, dass geringes Vertrauen, als Folge von wiederholtem Vertrauensbruch, zu einer niedrigen Investitionsquote führt, die wiederum das Entstehen neuer Unternehmen als auch von Beschäftigung ungünstig beeinflusst (z. B. Zak & Fakhar 2006). Die Folge ist, dass Länder mit niedrigem Vertrauensniveau im Regelfall relativ arme Länder sind und der Wohlstand dort somit verhältnismäßig niedrig ist (z. B. Zak & Knack 2001). Doch nicht nur auf volkswirtschaftlicher Ebene hat ein geringes Vertrauensniveau negative Konsequenzen. In einem aktuellen Beitrag zu Vertrauen aus psychologischer Perspektive analysieren Simpson und Vieth (2022) verschiedene Bereiche (z. B. Vertrauen in der Familie, in der Ehe sowie am Arbeitsplatz) und kommen nach Sichtung der Fachliteratur zum Schluss, dass Vertrauen viele positive Effekte hat. Hohes Vertrauen in einer Paarbeziehung kann beispielsweise den im Zuge einer Elternschaft erlebten Stress reduzieren und hohes Vertrauen zwischen den Mitgliedern von Projektteams wirkt sich positiv auf die Team-Leistung aus (Simpson & Vieth 2022). Vor dem Hintergrund solcher und vieler weiterer in ihrem Beitrag vorgestellten Befunde verwundert es nicht, dass Simpson und Vieth (2022) im Fazit unter anderem schreiben: „Being able to trust others is a fundamental prerequisite for establishing and maintining various types of stable, satisfying relationships“ (S. 31). In einer Gesamtschau von Forschungsbefunden aus verschiedensten Fachdisziplinen wie Psychologie, Soziologie, Managementforschung, Informationssystem-Forschung, Volkswirtschaftslehre und Philosophie kann festgestellt werden, dass Vertrauen auf den Ebenen des Individuums, der Organisation sowie der Gesellschaft und Volkswirtschaft vieles Positives bewirkt – mit anderen Worten, ohne Mindestmaß an Vertrauen gäbe es keine funktionierende Wirtschaft und Gesellschaft.

Neben den benannten Fachdisziplinen kann Vertrauen auch aus der Sicht der Evolutionspsychologie und -biologie betrachtet werden. Zwischenmenschliches Vertrauen kann als eine adaptive Eigenschaft angesehen werden, die sich im Laufe der Zeit entwickelt hat, um die Zusammenarbeit und den sozialen Zusammenhalt innerhalb von Gruppen zu verbessern (z. B. Axelrod 1984). Menschen, die in der Lage waren, anderen zu vertrauen und Allianzen zu bilden, konnten in der Umwelt vergangener Zeiten, in der die Zusammenarbeit und der Zusammenhalt für das Jagen, Sammeln und den Schutz vor Bedrohungen unerlässlich waren, eher überleben und sich erfolgreich fortpflanzen. Weiter wurde diese Tendenz zum Vertrauen möglicherweise durch positive Erfahrungen mit reziprokem Altruismus verstärkt, bei dem die Individuen vom gegenseitigen Austausch von Ressourcen oder durch gegenseitige Unterstützung profitierten (z. B. Berg et al. 1995, McCabe et al. 2003, Nowak & Sigmund 2005). Darüber hinaus könnte die Fähigkeit, die Vertrauenswürdigkeit anderer zu erkennen, einen Fortpflanzungsvorteil mit sich gebracht haben (Riedl et al. 2011). Diese Fähigkeit ermöglichte es den Menschen, selektiv Allianzen mit denjenigen zu bilden, die mit größerer Wahrscheinlichkeit Kooperation und Unterstützung durch reziprokes Verhalten erwidern würden. Zudem ist zu erwähnen, dass Neugeborene und Kinder bis zur Jugend auf andere Menschen angewiesen sind, um überleben zu können (Lieberman & Eisenberger 2009, Riedl & Javor 2012). Folglich ist das Überleben von Neugeborenen und Kindern und damit das Überleben der Menschheit im Allgemeinen von der sozialen Bindung zwischen Menschen abhängig, und eine solche Bindung setzt Vertrauen voraus. Im Einklang mit dieser Argumentation hat eine Studie von De Dreu et al. (2010) festgestellt, dass die intranasale Verabreichung des neuroaktiven Hormons Oxytocin das Vertrauen gegenüber Mitgliedern der eigenen Gruppe beeinflusst, nicht jedoch gegenüber Mitgliedern fremder Gruppen, mit denen man gegebenenfalls in Konkurrenz steht. Vor dem Hintergrund der vorgestellten Argumente und Befunde verwundert es nicht, dass in der Evolutionsbiologie auch die Rolle der genetischen Verwandtschaft bei der Förderung von Vertrauen hervorgehoben wird. Genetisch verwandte Individuen neigen eher dazu, sich zu vertrauen und untereinander zu kooperieren (z. B. Phelps et al. 2000). Das Vertrauen zwischen Menschen kann somit als ein adaptiver Mechanismus verstanden werden, der sowohl von der Evolutionspsychologie als auch von der Evolutionsbiologie geprägt wurde, um die Zusammenarbeit zu fördern, soziale Bindungen zu stärken und letztlich den Fortpflanzungserfolg zu erhöhen.

Aus Sicht der vorliegenden Dissertationsschrift sind insbesondere zwei Umstände von besonderer Relevanz, die sich aus einer Sichtung der aktuellen Vertrauensforschung (z. B. Krueger 2022) ergeben:

1. Zunehmend mehr werden die neurobiologischen Grundlagen von Vertrauen erforscht und damit einhergehend hat sich mittlerweile ein umfangreicher Fundus an empirischen Erkenntnissen entwickelt; dieser bezieht sich insbesondere auf drei Analyseebenen (Genetik, Hormone, Gehirn), die in Wechselwirkung miteinander stehen.
2. Individuelle Unterschiede (z. B. Geschlecht, Persönlichkeit) können einen signifikanten Einfluss auf die Vertrauenswahrnehmung und das Vertrauensverhalten haben, wobei diese Unterschiede – zumindest teilweise – auf neurobiologische Unterschiede zurückgeführt werden können.

In Anbetracht der beschriebenen Ausgangssituation befasst sich die vorliegende Dissertation mit

> *Vertrauen in einer digitalen Welt und berücksichtigt dabei die Neurobiologie des Vertrauens sowie ausgewählte individuelle Unterschiede (konkret Geschlecht, Persönlichkeit und exemplarisch für neuropsychopathologische Erkrankungen die Parkinson-Krankheit).*

Tabelle 1 fasst die dieser kumulativen Arbeit zugrunde liegenden Publikationen zusammen (siehe dazu auch Abbildung 1). Die Darstellung erfolgt in chronologisch aufsteigender Reihenfolge.

Nr.	*Referenz*	*Typ*
1	Riedl, R., Hubert, M., & Kenning, P. (2010). Are There Neural Gender Differences in Online Trust? An fMRI Study on the Perceived Trustworthiness of eBay Offers. *Management Information Systems Quarterly, 34*(2), 397-428.	E
2	Riedl, R., & Javor, A. (2012). The Biology of Trust: Integrating Evidence from Genetics, Endocrinology and Functional Brain Imaging. *Journal of Neuroscience, Psychology, and Economics, 5*(2), 63-91.	R
3	Riedl, R., Mohr, P., Kenning, P., Davis, F. D., & Heekeren, H. (2014). Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory. *Journal of Management Information Systems, 30*(4), 83-113.	E
4	Javor, A., Ransmayr, G., Struhal, W., & Riedl, R. (2016). Parkinson Patients' Initial Trust in Avatars: Theory and Evidence. *PLoS ONE, 11*, 1-21.	E
5	Riedl, R., Javor, A., Gefen, D., Felten, A., & Reuter, M. (2017). Oxytocin, Trust, and Trustworthiness: The Moderating Role of Music. *Journal of Neuroscience, Psychology, and Economics, 10*(1), 1-8.	E
6	Riedl, R. (2022b). Is Trust in Artificial Intelligence Systems Related to User Personality? Review of Empirical Evidence and Future Research Directions. *Electronic Markets, 32*, 2021-2051.	R
7	Riedl, R., Hogeterp, S., & Reuter, M. (2024). Do Patients Prefer a Human Doctor, Artificial Intelligence, or a Blend, and is This Preference Dependent on Medical Discipline? Empirical Evidence and Implications for Medical Practice. *Frontiers in Psychology, 15*, 1-13.	E

Tabelle 1: Publikationen der vorliegenden kumulativen Dissertationsschrift

Anmerkung: E = Empirischer Artikel, R = Review.

Es folgt eine Darstellung des Aufbaus dieser Dissertationsschrift. In Kapitel 2.1 wird zunächst die Frage beantwortet, wie Vertrauen in der Fachliteratur konzeptualisiert und gemessen wird. Danach wird in Kapitel 2.2 ein Modell eingeführt, um Vertrauen in einer digitalen Welt zu beschreiben. Das Modell dient im weiteren Verlauf auch dazu, die Publikationen der vorliegenden Arbeit konzeptionell einzuordnen. In Kapitel 2.3 werden die neurobiologischen Grundlagen von Vertrauen beschrieben – diese Darstellung ist eine kompakte Zusammenfassung von Publikation Nr. 2, wobei aktuelle Forschungserkenntnisse seit dem Erscheinen dieser Publikation berücksichtigt werden. In Kapitel 2.4 wird auf Vertrauen und individuelle Unterschiede eingegangen, wobei eine Gliederung in drei Sub-Kapitel erfolgt, und zwar entlang der in dieser Arbeit untersuchten Faktoren Geschlecht (Kapitel 2.4.1), Persönlichkeit (Kapitel 2.4.2) und Parkinson-Krankheit (Kapitel 2.4.3). In Kapitel 3 folgen

Zusammenfassungen der dieser Dissertation zugrunde liegenden Publikationen (siehe Tabelle 1). In Kapitel 4 erfolgt die Darstellung des Erkenntnisbeitrags der vorliegenden Dissertationsschrift, der in zwei Sub-Kapitel gegliedert ist, nämlich in Zusammenfassung, Implikationen und bisheriger wissenschaftlicher Impact (Kapitel 4.1) sowie Limitationen und zukünftige Forschung (Kapitel 4.2). In Kapitel 5 folgt das Literaturverzeichnis und das abschließende Kapitel 6 umfasst in einem Anhang alle Originalpublikationen, die der vorliegenden kumulativen Dissertationsschrift zugrunde liegen.

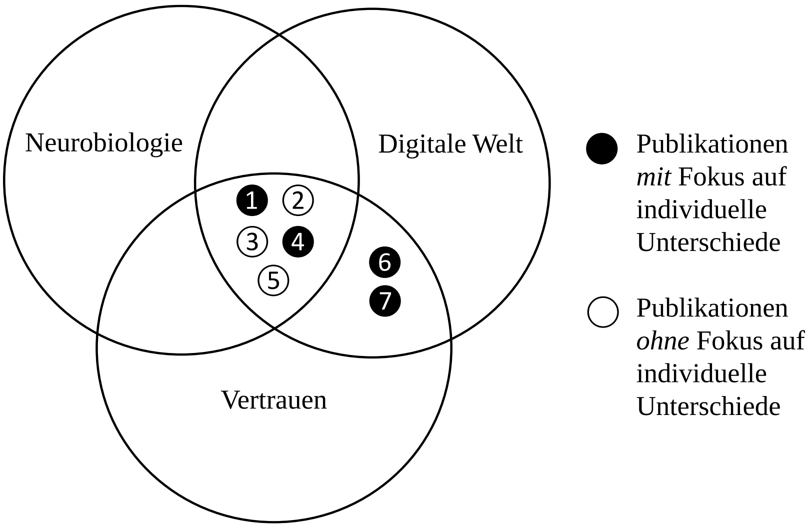


Abbildung 1: Einordnung der Publikationen der vorliegenden Dissertationsschrift

Anmerkung: Die Nummern in der Abbildung beziehen sich auf die linke Spalte in Tabelle 1.

Abbildung 1 ordnet die sieben Publikationen der Dissertationsschrift in einem Venn-Diagramm ein (die Nummern beziehen sich auf die linke Spalte in Tabelle 1). Dabei werden die drei Mengen *Digitale Welt*, *Vertrauen* und *Neurobiologie* konzeptualisiert. Zudem gibt der Farbcode (schwarz, weiß) an, ob eine Publikation explizit auf zumindest einen individuellen Unterschied (z. B. Geschlecht) fokussiert. Es wird transparent, dass fünf Publikationen in der Schnittmenge aller drei Mengen liegen, wobei zwei davon einen Fokus auf individuelle Unterschiede legen (Nr. 1, Nr. 4) und drei nicht (Nr. 2, Nr. 3, Nr. 5). Weiter ist zu sehen, dass zwei Publikationen in der Schnittmenge von *Vertrauen* und *Digitale Welt* liegen, wobei sich beide Publikationen mit individuellen Unterschieden befassen (Nr. 6, Nr. 7).

2. Vertrauen

Dieses Kapitel 2 hat den Zweck, wesentliche Grundlagen dieser Dissertation zu beschreiben, die für das Verständnis und die Einordnung der in Kapitel 3 dargestellten Publikationen notwendig sind. Zu Beginn wird die Frage geklärt, was Vertrauen ist. Zudem wird auf die damit in Verbindung stehende Frage eingegangen, wie Vertrauen gemessen werden kann. Es folgt die Darstellung eines Modells, um

Vertrauen in einer digitalen Welt zu beschreiben. Danach werden wesentliche neurobiologische Grundlagen von Vertrauen erläutert, um dann noch auf Vertrauen und individuelle Unterschiede einzugehen (Geschlecht, Persönlichkeit und Parkinson-Krankheit).

2.1 Konzeptualisierung und Messung von Vertrauen

Vor dem Hintergrund der Allgegenwärtigkeit von Vertrauen in Situationen menschlicher Interaktion, sowohl in traditionellen als auch in virtuellen Umgebungen, haben verschiedene Wissenschaftsdisziplinen eine Vielzahl an Forschungsbefunden über das Wesen des Vertrauens sowie seine Ursachen (z. B. Gesichtsausdrücke in persönlicher Interaktionen zwischen Menschen oder gut gestaltete Benutzeroberflächen bei computerbasierter Interaktionen) und Konsequenzen (z. B. Informationsweitergabe oder kooperatives Verhalten) untersucht. Psychologen, Soziologen, Philosophen, Ökonomen sowie Forschende im Management- und Informationssystembereich haben das Vertrauen sowohl theoretisch als auch empirisch untersucht (z. B. Gefen et al. 2008; Schilke et al. 2021; Seppänen et al. 2007; Simpson 2012; Swan et al. 1999).

Trotz der Vielzahl von Disziplinen, die das Phänomen untersucht haben, gibt es eine weitgehende Übereinstimmung in Bezug auf die Konzeptualisierung des Vertrauens. Die meisten Wissenschaftler beschreiben Vertrauen als ein Verhalten, durch das eine Partei, der Vertrauensgeber (Trustor), sich den Handlungen einer anderen Partei, dem Vertrauensnehmer (Trustee), aussetzt. Der Vertrauensgeber wird verwundbar, weil ihm im Fall eines Vertrauensbruchs durch den Vertrauensnehmer ein Schaden entsteht. Wie groß dieser Schaden ist, hängt von der Vertrauenssituation ab und reicht beispielsweise von einem wirtschaftlichen Verlust bis zum Extremfall des Verlusts des eigenen Lebens.

In einem aktuellen Artikel, der im *Annual Review of Organizational Psychology and Organizational Behavior* erschienen ist, werden in einer historischen Betrachtung der Vertrauensforschung die folgenden beiden Definitionen als miteinander im Einklang stehend und als besonders wichtig für die Entwicklung des Forschungsfelds dargestellt und mit „widely adopted by trust scholars“ charakterisiert (Dirks & de Jong 2022, S. 250) – Vertrauen ist demnach:

> *„the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (Mayer et al. 1995, S. 712).*

> *„a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (Rousseau et al. 1998, S. 395).*

Das Vertrauensverhalten des Vertrauensgebers wird somit von seiner Überzeugung über die Vertrauenswürdigkeit des Vertrauensnehmers beeinflusst, die wiederum seine Einstellung gegenüber dem Vertrauensnehmer und nachfolgende Verhaltensabsichten beeinflusst. Zudem wird das Vertrauensverhalten in einer bestimmten Situation auch vom allgemeinen Vertrauensniveau des Vertrauensgebers (auch als Vertrauensdisposition bezeichnet) sowie von den individuellen Risikopräferenzen beeinflusst (Riedl & Javor 2012).

Wichtige Merkmale eines Vertrauensnehmers sind Fähigkeit, Wohlwollen und Integrität (Mayer et al. 1995). Wenn ein Vertrauensgeber an die Vertrauenswürdigkeit eines Vertrauensnehmers glaubt, so nimmt er an, dass der Vertrauensnehmer (1) Kompetenzen hat, die für die Beziehung wichtig sind (Fähigkeit), (2) dem Vertrauensgeber aufrichtig wohlgesonnen ist und nicht von egoistischen Gewinnmotiven geleitet wird (Wohlwollen), und (3) Prinzipien folgt, die der Vertrauensgeber akzeptabel findet (Integrität). Die Bedeutung jedes Merkmals kann aufgrund innerer Zustände des Vertrauensgebers variieren. So kann beispielsweise eine hohe Risikowahrnehmung in Bezug auf Vertrauensmissbrauch und eine diesbezügliche Aversion („betrayal aversion") die Aufmerksamkeit auf das Wohlwollen anstatt auf die Fähigkeit lenken. Zudem beeinflussen Kontextfaktoren die Merkmalsbedeutung. Man denke hier beispielsweise an das Tandem-Fallschirmspringen. Hier wird die Fähigkeit des Instruktors mit Sicherheit von den meisten Menschen als wichtigstes Merkmal erachtet.

Abbildung 2 fasst die Struktur einer Vertrauenssituation grafisch zusammen. Eine Person, der Vertrauensgeber, der zum ersten Mal eine unbekannte Person, den Vertrauensnehmer, trifft, wird seine Überzeugung hinsichtlich der Vertrauenswürdigkeit der Person auf der Grundlage von Stimuli (z. B. Erscheinungsbild und Gesichtsausdruck des Vertrauensnehmers sowie seiner Selbstkontrolle (z. B. Adolphs et al. 2005; Righetti & Finkenauer 2011; Todorov 2008; Winston et al. 2002) bilden, was wiederum die Einstellung gegenüber der Person, nachfolgende Verhaltensabsichten und letztendlich tatsächliches Verhalten beeinflusst (z. B. Vertrauen/Misstrauen und Annäherung/Vermeidung). Darüber hinaus können die Vertrauensdisposition des Vertrauensgebers sowie seine Risikopräferenzen jeden Zusammenhang in dieser kausalen Kette moderieren (z. B. Fehr 2009).

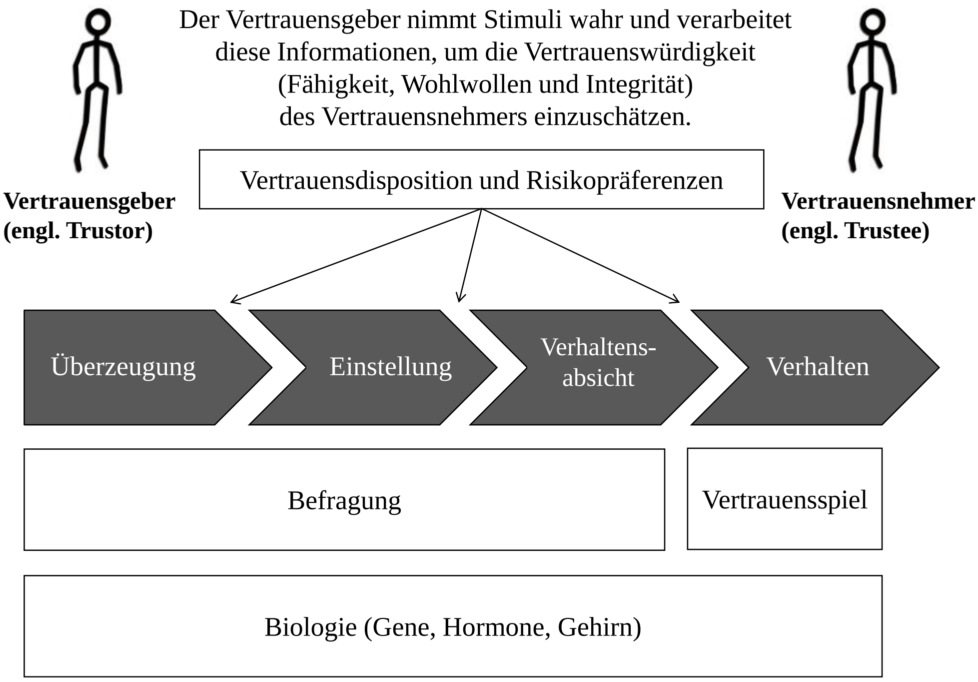


Abbildung 2: Struktur einer Vertrauenssituation (nach Riedl & Javor 2012, S. 64)

Obwohl die verhaltensorientierte Vertrauensforschung sowie Befragungsstudien eine Reihe von Erkenntnissen über das Wesen des Vertrauens sowie seine Ursachen und Konsequenzen offenbart haben (Rousseau et al. 1998; Seppänen et al. 2007; Swan et al. 1999), haben Wissenschaftler insbesondere in den letzten beiden Jahrzehnten das Thema auch aus einer neurobiologischen Perspektive untersucht. Diese neurobiologischen Untersuchungen lassen sich in drei Bereiche einteilen: Genetik (z. B. Cesarini et al. 2008), Endokrinologie (z. B. Kosfeld et al. 2005) und Gehirnfunktionalität (z. B. King-Casas et al. 2005). Es ist anzumerken, dass diese drei Bereiche nicht voneinander unabhängig sind – siehe dazu beispielsweise eine Studie von Baumgartner et al. (2008), in der die über einen Nasenspray verabreichte Wirkung des neuroaktiven Hormons Oxytocin auf Hirnaktivierungsmuster in Vertrauenssituationen untersucht wurde.

Ein Grund für diese Entwicklung hin zu einer neurobiologisch ausgerichteten Vertrauensforschung ist die Verfügbarkeit leistungsfähiger Methoden wie der funktionellen Magnetresonanztomographie (fMRT). Ein weiterer Grund ist die Erkenntnis, dass jegliches menschliches Verhalten, das zwischen Individuen variiert, zumindest teilweise mit neurobiologischen Faktoren in Zusammenhang steht (z. B. Cacioppo et al. 2000; Turkheimer 1998). So liegen die Erblichkeitsschätzungen für Vertrauen zwischen 10%-20% in Bezug auf Verhaltensdaten mit dem Vertrauensspiel (Cesarini et al. 2008) und bei 33 % (Männer) und 39 % (Frauen) in Bezug auf Selbstberichtsdaten (Oskarsson et al. 2012). Daraus folgt, dass die Berücksichtigung neurobiologischer Faktoren unerlässlich ist, um einen Teil der Varianz im menschlichen Vertrauensverhalten aufzuklären (z. B. Fehr & Camerer 2007).

Abbildung 2 zeigt weiter, dass die Messung von Vertrauen im Allgemeinen – unabhängig davon, ob die Forschung eine neurobiologische Komponente aufweist – einerseits durch Befragung (Selbstbericht) und andererseits durch das Vertrauensspiel (Trust Game; Verhaltensexperiment) erfolgt. Vertrauen wird oft als Überzeugung, Einstellung oder Verhaltensabsicht konzeptualisiert. Diese Konzeptualisierungen werden verwendet, um Befragungsinstrumente zur Messung von Vertrauen zu entwickeln. Wenn Vertrauen beispielsweise als Überzeugung konzeptualisiert wird, wird das Konstrukt anhand der Eigenschaften eines Vertrauensnehmers (z. B. Fähigkeit, Wohlwollen und Integrität) operationalisiert.[1] So lauten beispielsweise drei Items zur Messung von Wohlwollen auf einer Likert-Skala: (1) „Ich glaube, dass der Vertrauensnehmer in meinem besten Interesse handeln würde.“, (2) „Wenn ich Hilfe brauche, wird der Vertrauensnehmer sein Bestes tun, um mir zu helfen.“ und (3) „Der Vertrauensnehmer interessiert sich für mein Wohlergehen, nicht nur für das eigene.“ (wörtliche Übersetzung aus dem Englischen nach McKnight et al. 2002, S. 355). Wenn Vertrauen als Einstellung oder Verhaltensabsicht konzeptualisiert wird, werden ähnliche Items verwendet, um Vertrauen zu messen (McKnight et al. 2002).

In diesem Zusammenhang sei der weithin bekannte *World Values Survey* erwähnt, der unter anderem Fragen zur Vertrauenseinstellung enthält. Die zentrale Frage lautet (im englischen Original): „Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?” (Antwortmöglichkeiten: „1. Most people can be trusted.” und „2. Need to be very careful.”) (OECD 2017, S. 190). In Abbildung 3 werden beispielhaft entlang einer Längsschnittdarstellung acht Nationen dargestellt. Auf der Y-Achse ist der Prozentsatz jener Menschen dargestellt, die der Aussage „most people can be trusted“ zustimmen, also eine positive Vertrauenseinstellung haben (man beachte, dass für die dargestellten Nationen nicht für alle Jahre Daten vorhanden sind). Man sieht beispielsweise, dass aktuell in China rund zwei Drittel der Bevölkerung eine positive Vertrauenseinstellung haben, wohingegen es in Ägypten oder Kolumbien weniger als 10 % sind. In Deutschland hatten im Jahr 2022 rund 45 % der Menschen eine positive Vertrauenseinstellung. Die aktuellste Ausgabe des gleichfalls weithin bekannten und auch fragebogenbasierten *Edelman Trust Barometer* (https://www.edelman.com/trust/) kommt zu ähnlichen Ergebnissen. Beispielsweise wird für Deutschland in Bezug auf die Metrik „Trust Index“ (definiert als „the average percent trust in NGOs, business, government and media“, S. 40) ein Wert von 46 % angegeben; auch hier führt China das Ranking an (Edelman 2023).

[1] Der hier im Kontext von Befragungsstudien angesprochene Vertrauensnehmer entspricht dem Entscheider 2 (second mover) im Kontext von Verhaltensstudien auf der Basis des Vertrauensspiels (siehe Abbildung 4).

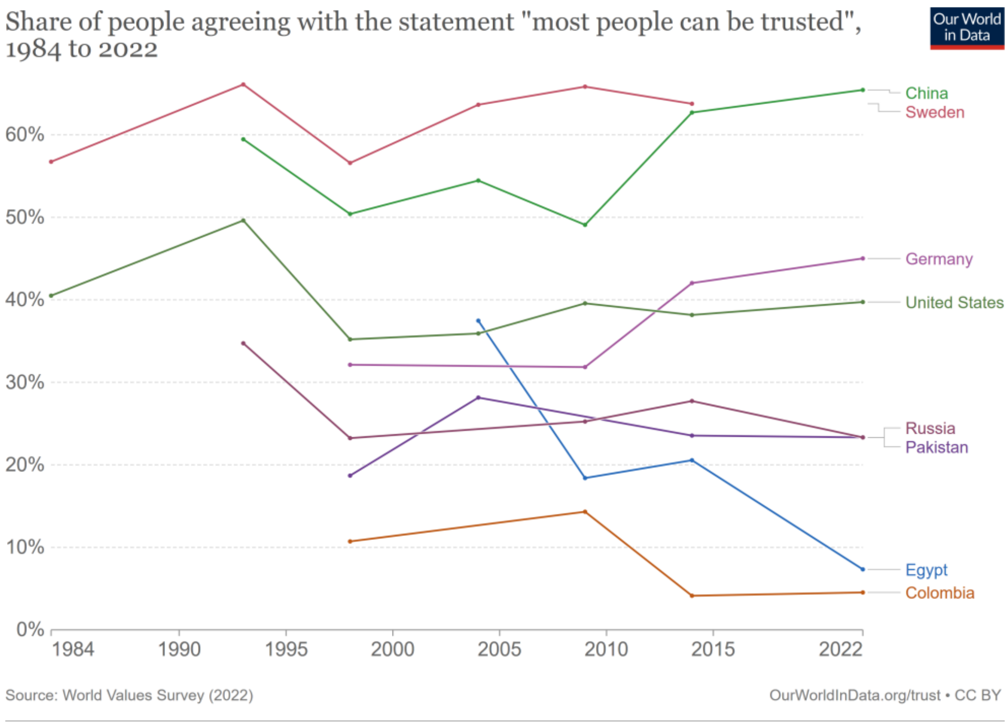


Abbildung 3: Anteil an Personen mit positiver Vertrauenseinstellung (Quelle: *World Values Survey* 2022, zitiert nach https://ourworldindata.org/trust)

Viele Studien in unterschiedlichen Wissenschaftsdisziplinen messen Vertrauen auf der Basis von Befragungsinstrumenten; in der Praxis ist dies ohnehin der Regelfall (z. B. *World Values Survey*). Daraus folgt, dass viele Studien kein tatsächliches Vertrauensverhalten untersuchen. Stattdessen werden Antezedenzien des Vertrauensverhaltens (z. B. Überzeugung hinsichtlich der Vertrauenswürdigkeit eines Vertrauensnehmers sowie diesbezügliche Einstellungen und Verhaltensabsichten) gemessen. Wenn Vertrauen jedoch als Verhalten konzeptualisiert wird, ist es unerlässlich, tatsächliches Verhalten (z. B. Investmentverhalten) in einer Vertrauenssituation (z. B. im Vertrauensspiel) zu messen, anstatt die Menschen auf der Basis von hypothetischen Vertrauenssituationen zu befragen. Ein wichtiger Unterschied zwischen dem Selbstbericht und Vertrauensspiel ist, dass das Verhalten bei letzterem mit einer realen Konsequenz verknüpft ist (meist dem Risiko, eigenes Geld zu verlieren, wenn sich der Vertrauensnehmer als nicht vertrauenswürdig herausstellt); Selbstberichte können hingegen von Antworttendenzen (z. B. positive Selbstdarstellung) verfälscht sein.[2]

Das Vertrauensspiel wurde entwickelt, um sowohl das Vertrauen als auch die Vertrauenswürdigkeit als tatsächliches Verhalten der Spieler in einem ökonomischen Spiel zu messen (Berg et al. 1995; Camerer & Weigelt 1988; McCabe et al. 2003; McCabe & Smith 2000). In einer gängigen Version des

[2] Diese Aussage soll den Erkenntniswert von Befragungsstudien nicht schmälern. Vielmehr geht es darum, auf den komplementären Erkenntniswert des Einsatzes von Verhaltensmessung hinzuführen.

Vertrauensspiels hat Entscheider 1 einen anfänglichen Geldbetrag von x Geldeinheiten (z. B. 10 €, siehe Abbildung 4). Zuerst entscheidet die Person, ob sie den Betrag teilen möchte (z. B. 5 € für jeden Spieler, dann endet das Spiel, siehe linker Pfad in Abbildung 4) oder ob sie (einen Teil davon) an Entscheider 2 senden möchte (rechter Pfad in Abbildung 4). Entscheider 2 beobachtet die Handlung von Entscheider 1 und entscheidet dann, ob er den Betrag behalten oder (einen Teil davon) mit Entscheider 1 teilen möchte. Der Versuchsleiter multipliziert den von Entscheider 1 gesendeten Geldbetrag (im Beispiel wird der Betrag verdreifacht), sodass beide Spieler gemeinsam mehr davon profitieren, wenn Entscheider 1 Geld übermittelt und Entscheider 2 einen ausreichenden Betrag zurückschickt. In dem in Abbildung 4 dargestellten Beispiel hat Entscheider 2 zwei Möglichkeiten: entweder das Geld zu teilen (d. h. jeder Spieler erhält 15 €) oder das gesamte Geld zu behalten (d. h. Entscheider 1 erhält 0 €, Entscheider 2 erhält 30 €).

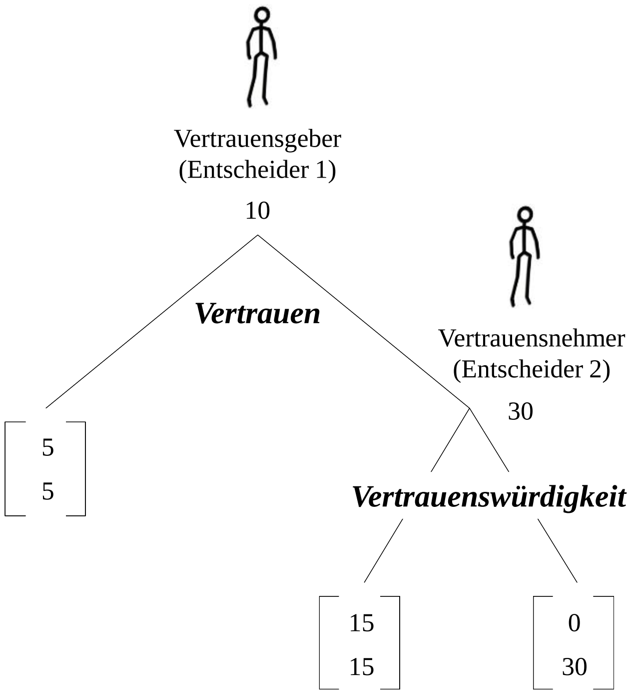


Abbildung 4: Struktur des Vertrauensspiels (nach Riedl & Javor 2012, S. 67)
Anmerkung: Der obere Wert in den eckigen Klammern gibt die Auszahlung an den Vertrauensgeber an (= Entscheider 1), der untere Wert die Auszahlung an den Vertrauensnehmer (= Entscheider 2). Die Geldbeträge und die Auszahlungsmatrix in der Abbildung sind beispielhaft, woraus folgt, dass es unterschiedliche Version des Vertrauensspiels gibt.

Das Vertrauensspiel simuliert einen sequentiell ablaufenden wirtschaftlichen Austausch in Abwesenheit von Institutionen zur Vertragsdurchsetzung (Fehr & Camerer 2007). Wenn das Spiel einmalig gespielt wird, hat Entscheider 2 einen starken Anreiz, das gesamte Geld zu behalten und nichts an Entscheider 1 zurückzuzahlen. Wenn Entscheider 1 jedoch dieses Verhalten antizipiert, gibt es wenig Grund für ihn, Geld zu übermitteln. Folglich würde, wenn Entscheider 1 kein Geld übermitteln würde, eine Chance für einen höheren gegenseitigen Profit verloren gehen. Im Vertrauensspiel sind der von Entscheider 1 gesendete Betrag als Verhaltensmaß für Vertrauen und die Rücküberweisung von Entscheider 2 als Verhaltensmaß für Vertrauenswürdigkeit konzeptualisiert (siehe Abbildung 4).

Basierend auf ökonomischen Standardannahmen hinsichtlich Rationalität und Egoismus lautet die Vorhersage für das Vertrauensspiel, dass Entscheider 1 kein Geld an Entscheider 2 übermittelt und folglich auch kein Geldbetrag zurückgegeben wird. Diese theoretische Vorhersage steht jedoch in Widerspruch zu empirischen Forschungsergebnissen. Eine Meta-Analyse mit 162 Studien und über 23.000 Teilnehmern zeigt folgendes: Entscheider 1 übermittelt in der Regel etwa die Hälfte des anfänglichen Betrags (0,502) an Entscheider 2 und dieser sendet häufig einen signifikanten Betrag zurück (0,372) (Johnson & Mislin 2011). Darüber hinaus zeigte diese Meta-Analyse systematische Unterschiede zwischen geografischen Regionen: Gesendeter Anteil (Vertrauen) – Europa (0,537), Nordamerika (0,517), Asien (0,482), Südamerika (0,458), Afrika (0,456); zurückgesendeter Anteil (Vertrauenswürdigkeit): Europa (0,382), Nordamerika (0,340), Asien (0,460), Südamerika (0,369), Afrika (0,319).

2.2 Vertrauen in einer digitalen Welt

Die vorliegende Dissertation fokussiert auf Vertrauen in einer digitalen Welt. Um das Wesen des Vertrauens in einer digitalen Welt besser zu verstehen, wurde ein Beschreibungsmodell entwickelt (Riedl 2022a). Dieses Modell umfasst vier Interaktionsszenarien:

- *Szenario A* beschreibt eine **technologiefreie Situation zwischenmenschlicher Interaktion** (z. B. Interaktion von Mitgliedern einer Familie, von Mitarbeitern in einem Unternehmen sowie von Verkäufer und Käufer von Angesicht zu Angesicht); dieses Szenario dient primär als Benchmark für die weiteren drei Szenarien.
- *Szenario B* beschreibt eine **Situation der computerbasierten zwischenmenschlichen Interaktion** – es interagieren somit zwei oder mehr Menschen über digitale Technologien (z. B. eine Verkäufer-Käufer-Interaktion auf eBay, Kommunikation über Social Media wie Facebook oder Instagram sowie zwischenmenschliche Interaktion über Partnervermittlungsplattformen wie Tinder oder Parship).
- *Szenario C* beschreibt eine **Situation der direkten Interaktion zwischen Mensch und Digitaltechnologie**, wobei die Manifestation der Technologie unterschiedliche Formen annehmen kann (z. B. Interaktion mit einem Roboter, einem autonomen Fahrzeug oder einem Check-in-Terminal am Flughafen).

- *Szenario D* beschreibt eine **Situation der computerbasierten Interaktion zwischen Mensch und Digitaltechnologie** (z. B. Kommunikation mit KI-Systemen wie ChatGPT oder Interaktion mit virtuellen Verkäufern auf Internetplattformen); man beachte, dass „computerbasiert" im Regelfall Interaktion über einen Personal Computer (PC), Laptop oder über ein Smartphone bedeutet.

Nr.	*Referenz*	*Szenarien*			
		A	*B*	*C*	*D*
1	Riedl, R., Hubert, M., & Kenning, P. (2010). Are There Neural Gender Differences in Online Trust? An fMRI Study on the Perceived Trustworthiness of eBay Offers. *Management Information Systems Quarterly, 34*(2), 397-428.		●		
2	Riedl, R., & Javor, A. (2012). The Biology of Trust: Integrating Evidence from Genetics, Endocrinology and Functional Brain Imaging. *Journal of Neuroscience, Psychology, and Economics, 5*(2), 63-91.	●	●		
3	Riedl, R., Mohr, P., Kenning, P., Davis, F. D., & Heekeren, H. (2014). Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory. *Journal of Management Information Systems, 30*(4), 83-113.				●
4	Javor, A., Ransmayr, G., Struhal, W., & Riedl, R. (2016). Parkinson Patients' Initial Trust in Avatars: Theory and Evidence. *PLoS ONE, 11*, 1-21.				●
5	Riedl, R., Javor, A., Gefen, D., Felten, A., & Reuter, M. (2017). Oxytocin, Trust, and Trustworthiness: The Moderating Role of Music. *Journal of Neuroscience, Psychology, and Economics, 10*(1), 1-8.	●	●		
6	Riedl, R. (2022b). Is Trust in Artificial Intelligence Systems Related to User Personality? Review of Empirical Evidence and Future Research Directions. *Electronic Markets, 32*, 2021-2051.			●	●
7	Riedl, R., Hogeterp, S., & Reuter, M. (2024). Do Patients Prefer a Human Doctor, Artificial Intelligence, or a Blend, and is This Preference Dependent on Medical Discipline? Empirical Evidence and Implications for Medical Practice. *Frontiers in Psychology*, *15*, 1-13.	●		●	●

Tabelle 2: Inhaltlicher Fokus der Publikationen dieser Dissertationsschrift entlang der vier Szenarien

Das gemeinsame Merkmal aller Szenarien ist, dass ein Mensch in der Rolle des Vertrauensgebers handelt, während der Vertrauensnehmer entweder ein anderer Mensch (Szenario A und B) oder ein technologisches Artefakt (Szenario C und D) ist. Wie bereits ausgeführt liegt der Fokus der vorliegenden Dissertation auf der Untersuchung von Phänomenen in einer digitalen Welt. Tabelle 2 fasst den inhaltlichen Fokus der Publikationen dieser Dissertationsschrift entlang der vier Szenarien zusammen. In einer Gesamtschau wird transparent, dass die einzelnen Publikationen alle vier Szenarien adressieren.

In der Fachliteratur gibt es hinsichtlich der Szenarien C und D eine Debatte darüber, ob es sinnhaft ist, Digitaltechnologien (z. B. KI-Systeme oder Maschinen) als Vertrauensnehmer zu betrachten (z. B. Wang & Benbasat 2005). Diejenigen, die argumentieren, dass eine Digitaltechnologie kein

Vertrauensnehmer sein kann, nehmen an, dass Technologien wichtige menschliche Eigenschaften fehlen, insbesondere *Bewusstsein* sowie das *Gefühl des Verrats*, wenn Vertrauen gebrochen wird (z. B. Friedman et al. 2000; Friedman & Millett 1997). Aus dieser Sichtweise folgt, dass Digitaltechnologien – per Definition – nicht vertraut werden kann, sondern man kann sich auf sie verlassen oder auch nicht. Nach dieser Ansicht ist es nicht sinnvoll, Vertrauen überhaupt als Phänomen in der Mensch-Maschine-Interaktion zu untersuchen.

Im scharfen Gegensatz zu dieser Ansicht zeigen jedoch Forschungsarbeiten, die auf dem Computers Are Social Actors (CASA)-Paradigma[3] basieren, dass Menschen in ihrer Interaktion mit Digitaltechnologien ähnliche soziale Regeln und Heuristiken anwenden wie bei ihrer Interaktion mit Menschen (Nass et al. 1993; Nass et al. 1994; Reeves & Nass 1996). Zum Beispiel ist *Reziprozität* ein Phänomen, das direkt mit Vertrauen zusammenhängt (z. B. Krueger & Meyer-Lindenberg 2019; Riedl et al. 2011). Menschen wenden soziale Reziprozitätsregeln jedoch auch in ihrer Interaktion mit Computern an (Fogg & Nass 1997). Dementsprechend kann eine Digitaltechnologie ohne Bewusstsein ein Vertrauensnehmer sein, obwohl ihr menschliche Attribute wie *Intentionalität* oder *freier Wille* fehlen. Die herrschende Lehrmeinung ist somit, dass Digitaltechnologien jedenfalls Vertrauensnehmer sein können. Eine daraus folgende Frage ist, ob Menschen bei ihrer Interaktion mit technologischen Artefakten dieselben Vertrauenswürdigkeitsdimensionen anwenden wie in der zwischenmenschlichen Interaktion, nämlich Fähigkeit, Wohlwollen und Integrität. In diesem Zusammenhang schreiben Wang und Benbasat (2005, S. 76):

> *„[W]hile it may at first appear debatable that technological artifacts can be objects of trust, and that people assign human properties to them, evidence from a variety of relevant literature supports this argument. People respond socially to technological artifacts and perceive that they possess human characteristics (e.g., motivation, integrity, and personality). In particular, research findings have demonstrated that components of trust in humans and in technological artifacts do not differ significantly. This indicates that people not only utilize technological artifacts as tools, but also form social and trusting relationships with them."*

[3] Das vom US-amerikanischen Wissenschaftler Clifford Nass begründete CASA-Paradigma befasst sich primär mit dem Phänomen des Anthropomorphismus. Hierbei schreiben Menschen Technologien wie Computern menschenähnliche Eigenschaften und soziale Verhaltensweisen zu. Es werden einerseits die kognitiven Prozesse, die hinter solchen Zuschreibungen stehen, untersucht, und andererseits befasst man sich mit den komplexen psychologischen und verhaltensbezogenen Auswirkungen solcher Zuschreibungen, insbesondere jenen, die sich auf die Interaktion und Wahrnehmung der Benutzer beziehen. Durch die Untersuchung des Phänomens, wie Individuen nicht-menschlichen Entitäten wie Computern Persönlichkeitsmerkmale, Emotionen und soziale Verhaltensnormen zuschreiben, beleuchtet das CASA-Paradigma das komplexe Zusammenspiel zwischen menschlicher Kognition, Technologie und sozialer Dynamik; weiter bietet es Einblicke in die sich entwickelnde Beziehung zwischen Menschen und Maschinen in einer von Digitalisierung geprägten Gesellschaft.

Angesichts dieser Aussage und weiterer empirischer Befunde aus dem CASA-Paradigma (z. B. Brave et al. 2005) scheint es – auf den ersten Blick zumindest – zweckmäßig zu sein, dass Menschen dieselben Vertrauensdimensionen anwenden, obwohl es tatsächlich nicht rational ist, ein technologisches Artefakt ohne die menschlichen Eigenschaften Bewusstsein, Gefühl des Verrats, Intentionalität und freier Wille beispielsweise als wohlwollend zu betrachten.

Untersuchungen zeigen, dass sich die Vertrauensdimensionen ändern, je nachdem, ob sich das Vertrauen auf einen Menschen oder eine Maschine richtet (z. B. Hancock et al. 2011; McKnight et al. 2011). Daraus folgt, dass es zweckmäßig ist, zwischen Vertrauensdimensionen der Szenarien A und B (Vertrauensnehmer ist ein Mensch) und Vertrauensdimensionen der Szenarien C und D (Vertrauensnehmer ist eine Digitaltechnologie) zu unterscheiden. Wie in Riedl (2022a) im Detail ausgeführt, finden in den Szenarien A und B die Dimensionen Fähigkeit, Wohlwollen und Integrität Anwendung, wohingegen in den Szenarien C und D Leistung (anstatt Fähigkeit), Hilfestellung (anstatt Wohlwollen) und Berechenbarkeit (anstatt Integrität) Berücksichtigung finden. Mit anderen Worten, wenn ein Mensch über die Vertrauenswürdigkeit einer Digitaltechnologie (Maschine) befindet, dann wird dies unter anderem auf der Basis von Überlegungen zu Leistung, Hilfestellung und Berechenbarkeit erfolgen.[4] Man denke hier beispielsweise an die Verwendung von ChatGPT oder von einem Check-in-Terminal am Flughafen.

2.3 Neurobiologische Grundlagen von Vertrauen

Die Frage, wie unsere Gene einerseits und Umweltfaktoren[5] andererseits zur Manifestation menschlichen Verhaltens (z. B. Vertrauensverhalten) beitragen, ist insbesondere in der Psychologie eine grundlegende Forschungsfrage (im internationalen Schrifttum unter der Debattenbezeichnung *Nature vs. Nurture* bekannt). Obwohl es – zumindest in der Vergangenheit – Wissenschaftler gab, die einen absolut dominanten Einfluss von Genen einerseits sowie von Umwelteinflüssen andererseits vermuteten, sind sich heute die meisten Wissenschaftler darüber einig, dass beide Faktoren wichtig sind und keiner von beiden menschliches Verhalten in deterministischer Weise beeinflusst (Johnson 2007). Empirische Evidenz untermauert die Vorstellung, dass menschliches Verhalten das Ergebnis des komplexen Zusammenspiels von sowohl genetischer als auch von Umweltfaktoren ist (z. B. Bouchard 1994; Cacioppo et al. 2000). Auf der Basis dieser allgemeinen Feststellung wird im Folgenden der Forschungsstand zur Neurobiologie des Vertrauens zusammengefasst.

---

[4] In Riedl (2022a, S. 65) werden auf der Basis mehrerer Quellen folgende Definitionen angegeben: Leistung (performance): „the belief about the capability of the technology to accomplish its designated purpose”, Hilfestellung (helpfulness): „the belief that the specific technology provides adequate and responsive help for users”, Berechenbarkeit (predictability): „the belief that the technology will do what it is claimed to do without adding anything malicious on top of it”.

[5] Der Begriff „Umweltfaktoren“ ist hier breit zu verstehen und umfasst somit Faktoren wie Erfahrungen in der Kindheit, Sozialisierung im Allgemeinen sowie kulturelle Faktoren.

Zu Beginn wird auf die Hirnforschung eingegangen, gefolgt von Forschung zu Hormonen. Am Ende folgt noch eine kurze Zusammenfassung wesentlicher Erkenntnisse zu Vertrauen aus genetischer Perspektive, obwohl diese Analyseebene im Rahmen der vorliegenden Dissertation empirisch nicht untersucht wurde, sondern ein Fokus auf die Hirnebene (auf der Basis von fMRT sowie der Pathologie der Parkinson-Krankheit) und das neuroaktive Hormon Oxytocin gelegt wurde (siehe Tabelle 1).

Eine von Krueger und Meyer-Lindenberg (2019) veröffentlichte Arbeit beschreibt ein Modell zum zwischenmenschlichen Vertrauen („Model of Interpersonal Trust“), das auf der Basis von Erkenntnissen aus den Neurowissenschaften, der Psychologie und der Ökonomie entwickelt wurde. Zur Zielsetzung dieses Beitrags steht zu lesen: „[W]e present a neuropsychoeconomic (NPE) model of interpersonal trust, which provides a more integrative picture compared with previous relatively descriptive neuroscience reviews”. Krueger und Meyer-Lindenberg referenzieren vier existierende Reviews, darunter auch die Arbeit von Riedl und Javor (2012), die Teil dieser Dissertationsschrift ist. Nach Ansicht des Verfassers dieser Dissertation vollbringt der Beitrag von Krueger und Meyer-Lindenberg eine bemerkenswerte Integrationsleistung und fasst eine Vielzahl von Einzelbefunden zu einem kohärenten Bild zusammen, insbesondere werden Befunde aus der Hirnforschung berücksichtigt. Das Modell ist in Abbildung 5 dargestellt.

Teil (A) dieser Abbildung zeigt psychologische Konzepte, die in Vertrauenssituationen bedeutsam sind, wobei die Grundannahme getroffen wird, dass Vertrauen in Reziprozität (Verhalten, grau) ein soziales Dilemma repräsentiert, das den Entschluss eines Vertrauensgebers umfasst, sich dem Risiko, verraten zu werden, auszusetzen (Gefühle, rot). Dieses „Sich-einem-Risiko-aussetzen“ basiert auf der Erwartung (Kognition, blau), dass die reziproke Handlung eines Vertrauensnehmers mit einer Belohnung einhergeht (Motivation, grün). Teil (B) visualisiert die Entstehung von Vertrauen und die Vertrauenskomponenten. Vertrauen entsteht durch das Zusammenspiel der T-R-U-S-T-Komponenten (treachery, reward, uncertainty, strategy, trustworthiness), die mit Motivation, Gefühlen und Kognition in Beziehung stehen. Teil (C) zeigt, dass die Modellkomponenten neuronale Korrelate in domänenübergreifenden Hirnnetzwerken haben (man beachte hier den verwendeten Farbcode). Die Erwartung einer Belohnung in einer Vertrauenssituation (grünes Rechteck, Motivation, RWN = Reward Network) wird mit dem Risiko, verraten zu werden (rotes Rechteck, Gefühle, SAN = Salience Network), kontrastiert, was Unsicherheit (lila Ellipse) erzeugt. Um die Unsicherheit zu beseitigen, können zwei Arten von begrenzter Rationalität (Kognition) eingesetzt werden, bei denen SAN als Schalter fungiert und entweder das CEN = Central-Executive Network (extern gerichtete Kognition) oder das DMN = Default-Mode Network (intern gerichtete Kognition) aktiviert. Vertrauensgeber können eine Strategie auf der Basis extrinsischer Anreize (dunkelblaues Rechteck; kognitive Kontrolle) anwenden, um wirtschaftliche Vorteile zu erzielen. Es wird die Unsicherheit reduziert, weil das Verratsrisiko in eine positive wirtschaftliche Erwartung zur Reziprozität transformiert wird (durchgezogene Linien in Teil B). Vertrauensgeber können auch eine Strategie auf der Basis intrinsischer Anreize anwenden. Hierbei wird die beziehungsbezogene Vertrauenswürdigkeit bewertet (hellblaues Rechteck; soziale Kognition, DMN), um zum Erfolg der Beziehung beizutragen (soziale Rationalität) und um dadurch die Unsicherheit zu reduzieren; es wird das soziale Risiko eines Verrats in sozial positive Erwartungen der Reziprozität transformiert (gestrichelte Linien in Teil B). Ergänzend zur Arbeit von Krueger und Meyer-Lindenberg (2019) zeigt eine Untersuchung von Haas et al. (2015), dass die *Vertrauensneigung* mit strukturellen Unterschieden im Gehirn in Beziehung steht, insbesondere wird über Unterschiede in der grauen Substanz in folgenden Arealen berichtet: vmPFC, Amygdala und vordere Insula. Diese Ergebnisse stehen im Einklang mit dem Modell von Krueger und Meyer-Lindenberg (2019).

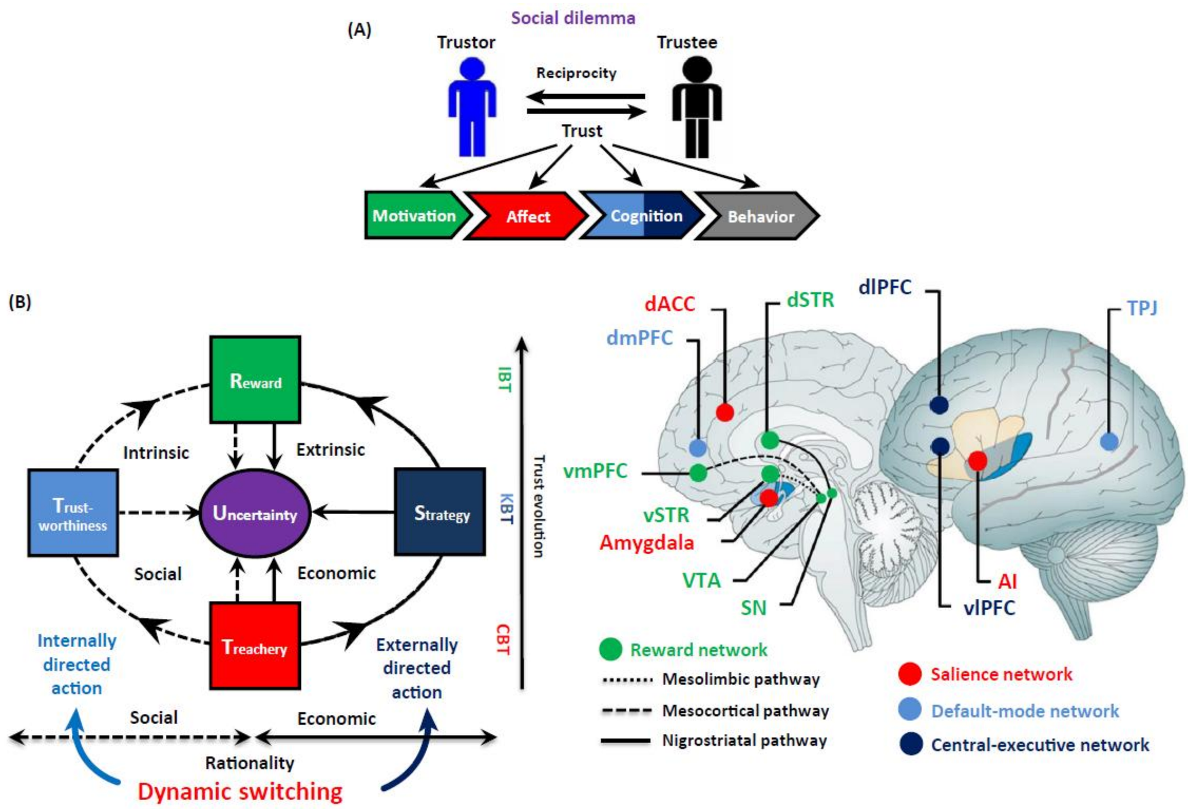


Abbildung 5: Neuropsychoökonomisches Modell zum zwischenmenschlichen Vertrauen (Quelle: Krueger und Meyer-Lindenberg 2019, S. 96)[6]

[6] AI = anterior insula; dACC = dorsal anterior cingulate cortex; dlPFC = dorsolateral prefrontal cortex; dmPFC = dorsomedial prefrontal cortex; dSTR = dorsal striatum; SN = substantia nigra; TPJ = temporoparietal junction; vlPFC = ventrolateral prefrontal cortex; vmPFC = ventromedial prefrontal cortex; vSTR = ventral striatum; VTA = ventral tegmentum area.

Vertrauenseinstellungen sowie -verhalten stehen auch mit Hormonen in Zusammenhang. Riedl und Javor (2012) haben einen Review vorgelegt und sieben Hormone identifiziert, die mit Vertrauen bzw. Misstrauen einen – mehr oder weniger direkten – Zusammenhang aufweisen. Sie gliedern die Hormone in ihrer Ergebnisdarstellung in solche, die mit Annäherungsverhalten und somit mit Vertrauen in Zusammenhang stehen (Oxytocin, Östrogen, Dopamin, Serotonin) und in solche, die mit Vermeidungsverhalten und somit mit Misstrauen in Zusammenhang stehen (Arginin-Vasopressin, Kortisol, Testosteron). Ein weiterer Befund ist, dass Oxytocin im Vertrauenskontext auffällig oft untersucht wurde (z. B. Kosfeld et al. 2005). Dies verwundert nicht, ist doch seit langer Zeit nicht nur bekannt, dass Oxytocin für den Gebär- und Stillprozess eine zentrale Bedeutung hat, sondern auch in Bezug auf mütterliches Versorgungsverhalten sowie für die Anbahnung und Erhaltung sozialer Beziehungen im Allgemeinen wichtig ist (z. B. Donaldson & Young 2008; Heinrichs & Domes 2008; Kovács 1986).[7]

Bis ins letzte Jahrzehnt konnte man auf der Basis vorliegender Forschungsbefunde davon ausgehen, dass Oxytocin direkt vertrauenserhöhend wirkt. Zudem wurde über weitere Effekte einer exogenen Oxytocin-Verabreichung berichtet, die mit Vertrauen in Zusammenhang stehen könnten – unter anderem verbesserte Emotionserkennung sowie ansteigender Fokus auf die Augenregion beim Interaktionspartner sowie Reduktion von sozialem Stress, was Misstrauen reduzieren kann (z. B. Kirsch et al. 2005; Meyer-Lindenberg et al. 2011).

Insbesondere die Erkenntnis, dass Oxytocin direkt vertrauenserhöhend wirkt, wurde jedoch in der jüngeren Vergangenheit zunehmend in Frage gestellt (siehe dazu beispielsweise eine präregistrierte Replikationsstudie von Declerck et al. 2020). Nach aktuellem Forschungsstand ist somit davon auszugehen, dass unterschiedliche Faktoren einen Einfluss darauf haben, ob und wie stark Oxytocin (OXT) die Vertrauenseinstellung sowie das Vertrauensverhalten beeinflusst. Yan und Kirsch (2022) haben dazu kürzlich einen Review vorgelegt, dessen wesentliche Erkenntnis sie wie folgt zusammenfassen:

> *„Taking the evidence reported […] although the association between OXT and trust has been widely investigated during the past decades, a direct trust promoting effect of OXT as proposed by early studies has not been found. Instead, there is evidence that the relationship between OXT and trust is modulated by manifold factors which are closely interrelated, e.g., social affiliation, personality traits, gender, mental disorders, and genetic disposition […] these contextual factors influence basic social mechanisms*

[7] Oxytocin wird landläufig auch immer wieder als „Vertrauenshormon" bezeichnet, wobei auch immer wieder die Bezeichnungen „Bindungshormon" oder „Kuschelhormon" anzutreffen sind.

like anxiety reduction, social salience and affiliative motivation which themselves contribute to a more general social cognitive and prosocial effect of OXT. While the factors which have been described to influence the effect of OXT on trust are very similar to those found for other social domains, one could argue that there is no specific trust promoting effect of OXT but, depending on the social situation and the individual dispositions, the prosocial effect of OXT is also reflected in increased trusting behavior or the intention to trust" (S. 329-330).

Diese jüngsten Forschungserkenntnisse zu den Effekten von Oxytocin sind insbesondere bei der Interpretation der Ergebnisse von Riedl et al. (2017) zu berücksichtigen, Publikation Nr. 5 dieser Dissertationsschrift. Wie in Kapitel 3 noch näher ausgeführt wird, wurde in dieser Untersuchung auf der Basis von Blutproben der Oxytocin-Spiegel von Probanden mit Vertrauensvariablen (auf der Basis des Vertrauensspiels) korreliert, wobei zusätzlich untersucht wurde, ob Musik als Stimulus den Oxytocin-Spiegel sowie die Vertrauensvariablen beeinflusst.

Ein dritter Bereich der neurobiologischen Vertrauensforschung befasst sich mit Genetik. Bevor auf molekulargenetische Studien eingegangen wird, werden die Befunde von Zwillingsstudien beschrieben. Bereits vor rund 15 Jahren wurde gezeigt, dass Vertrauens*verhalten* zu 10-20 % genetisch determiniert ist, und ob jemand vertrauenswürdig ist, ist zu 17-18 % genetisch vorbestimmt (Cesarini et al. 2008; in der Untersuchung kam das Vertrauensspiel zur Anwendung). Vertrauens*einstellungen* sind im Gegensatz zum Verhalten nach aktuellem Forschungsstand signifikant höher genetisch vorbestimmt: Hiraishi et al. (2008): 31 %, Sturgis et al. (2010): 66 %, Oskarrson et al. (2012): 33 % (Männer) und 39 % (Frauen). Weiter bedeutsam ist ein Befund, der zeigt, dass Vertrauen vererbbar ist, nicht jedoch Misstrauen (Reimann et al. 2017).

In einem aktuellem Artikel fassen Shou et al. (2022) die Forschung zu Vertrauen und Genetik zusammen. Ausgehend vom Umstand, dass in der Forschung bereits bestimmte Hormone und Neurotransmitter mit Vertrauen in Zusammenhang gebracht wurden, verwundert es nicht, dass die folgenden vier Gene einen Zusammenhang mit Vertrauenseinstellungen und -verhalten aufweisen: Oxytocin Rezeptor Gen (OXTR), Arginine Vasopressin Rezeptor 1A Gen (AVPR1A), Dopamin D4 Rezeptor Gen (DRD4) und Serotonin Transporter Gen (SLC6A4). Da Oxytocin in Publikation Nr. 5 der vorliegenden Dissertation untersucht wurde, soll auf wesentliche Befunde zu OXTR kurz eingegangen werden – bezüglich dem aktuellen Forschungsstand zu AVPR1A, DRD4 und SLC6A4 wird auf die Synopse in Shou et al. (2022) verwiesen.

Bezüglich OXTR wurde in einer Studie herausgefunden, dass die Vertrauenseinstellung als Mediator zwischen der Punktmutation (engl. single nucleotide polymorphism, SNP) OXTR rs53576 und Vertrauensverhalten fungiert und dass dieser Effekt nur bei Männern, nicht jedoch bei Frauen zu beobachten ist, was die Autoren mit geschlechterspezifischen Unterschieden im Östrogenspiegel erklären (Nishina et al. 2015). Weiter wird darüber berichtet, dass OXTR rs53576 mit der Struktur und Funktion der Amygdala in Zusammenhang steht (Tost et al. 2010; Wang et al. 2014)[8]; dieser spielt – wie weiter oben bereits dargestellt wurde – in Vertrauenssituation eine wichtige Rolle einnimmt; siehe das Modell von Krueger und Meyer-Lindenberg (2019) in Abbildung 5.[9] Schließlich wurde in einer Studie herausgefunden, dass OXTR rs53576 die Beziehung von Erfahrungen in der Kindheit und späteren Vertrauenseinstellungen moderiert (Zheng et al. 2020). Diese Ergebnisse bringen die hohe Komplexität der Thematik aufgrund des Zusammenspiels von Umweltfaktoren und spezifischen Genvarianten sowie den daraus folgenden Einfluss auf das Vertrauen zum Ausdruck.

Es ist wichtig zu betonen, dass soziale Umgebungsfaktoren, die insbesondere während der Kindheit und Jugendzeit auf ein Individuum einwirken, maßgeblichen Einfluss auf die Vertrauenseinstellung und das Vertrauensverhalten im späteren Leben haben – ein solcher Befund schwächt den direkten Einfluss der Gene auf Vertrauen ab. Van Lange et al. (2014) argumentieren, dass Vertrauen hauptsächlich durch zwischenmenschliche Prozesse entwickelt und aufrechterhalten wird, die *kognitive Schemata* verstärken oder abschwächen. Insbesondere die Lebensumstände können jedoch auch die Entwicklung von Vertrauen prägen. Unter anderem wurde der *sozioökonomische Status* in der Kindheit mit Vertrauensniveaus im Erwachsenenalter korreliert (Stamos et al. 2019); die Ergebnisse zeigen, dass Menschen, die in Umgebungen mit niedrigerem Status aufwachsen, mehr Herausforderungen und Bedrohungen erleben, was sie typischerweise aufmerksamer für ihre Umgebung macht und sie gegenüber anderen Menschen misstrauischer werden lässt. Diese und weitere Studien (z. B. Graafland & Lous 2019) zeigen, dass sowohl die Umgebungen, in denen Personen aufgewachsen sind, als auch diejenigen, in denen sie derzeit leben, die Entwicklung des Vertrauens in andere Menschen prägen können, insbesondere in Fremde. Diese Erkenntnis zeigt, dass Vertrauen ein komplexes psychologisches Konstrukt ist und es daher unwahrscheinlich ist, dass ein direkter Zusammenhang zwischen Vertrauenseinstellungen und -verhalten durch einige wenige Genausprägungen erklärt werden kann.

[8] In diesem Zusammenhang ist eine Studie von Kirsch et al. (2005) bedeutsam; es wurde gezeigt, dass die Funktion der Amygdala durch Oxytocin moduliert wird. Konkret zeigt die Studie, dass Oxytocin die Aktivierung der Amygdala verringert und die Kopplung der Amygdala an Hirnstammregionen reduziert, die an autonomen und verhaltensbezogenen Manifestationen von Angst beteiligt sind.

[9] Es ist zudem bekannt, dass Oxytocin einen Beitrag zur Aufklärung von Wahrnehmungs- und Verhaltensvarianz bei Autismus leisten kann (z. B. Uzefovsky et al. 2019).

2.4 Vertrauen und individuelle Unterschiede

Individuelle Unterschiede beeinflussen das Denken, Fühlen und Handeln von Menschen, so auch in Vertrauenssituationen. Im Folgenden werden drei individuelle Unterschiede in Bezug auf Vertrauen erläutert, da diese in der vorliegenden Dissertation untersucht wurden.[10]

2.4.1 Vertrauen und geschlechterspezifische Unterschiede

Ergebnisse von Befragungsstudien als auch experimentelle Befunde zeigen, dass Männer im Allgemeinen mehr vertrauen als Frauen. Beispielsweise zeigen Befragungen, dass Frauen im Vergleich zu Männern Aussagen der Art „die meisten Menschen sind vertrauenswürdig“ weniger oft zustimmen (Alesina & La Ferrara 2002; Glaeser et al. 2000; Terrell & Barrett 1979). Eine Reihe von experimentellen Untersuchungen – meist auf der Basis des Vertrauensspiels – untermauern diese Ergebnisse und kommen daher zu dem Schluss, dass Männer mehr vertrauen als Frauen (Buchan et al. 2008; Snijders & Keren 1999). Eine kürzlich veröffentlichte Meta-Analyse, in der 94 Studien auf der Basis des Vertrauensspiels integriert wurden (N = 17.082 verschiedene Probanden aus 23 Nationen), bestätigt, dass Männer in der Rolle des Vertrauensgebers den Spielpartnern mehr vertrauen und daher im Durchschnitt höhere Beträge an den Vertrauensnehmer senden als Frauen (Van den Akker et al. 2020). In der Meta-Analyse wird zudem die folgende Erklärung für dieses Ergebnis auf der Basis der *Evolutionären Psychologie* dargeboten:

> *„These results suggest that men send more money than women do when their money is going to be multiplied, thereby creating an efficiency gain. This so-called „male multiplier effect" may be explained by a stronger psychological tendency in men to acquire resources […] In short, it pays for men to be relatively more risk-taking in contexts where they can acquire resources like status, goods and money. This is evidenced by many studies that find men to take more (social) risks than women across different ages and cultures, and from modern to traditional societies" (S. 1-2).*[11]

Eine weitere wichtige Frage in diesem Zusammenhang ist, ob ein Geschlecht vertrauenswürdiger ist als das andere. Institutionelle Regelungen deuten darauf hin, dass Frauen im Allgemeinen als

[10] Es sei erwähnt, dass in den Publikationen der vorliegenden Dissertation noch viele weitere Faktoren untersucht wurden, die individuelle Unterschiede darstellen (im Regelfall als Kontrollvariablen), diese hatten jedoch zum größten Teil keinen wesentlichen Einfluss auf gefundene Haupteffekte – siehe dazu beispielsweise Publikation Nr. 7 (Tabelle 2).

[11] In Vertrauensexperimenten wird oftmals auch das von sozialer Interaktion unabhängige Risikoverhalten der Probanden gemessen, um gefundene Effekte auch tatsächlich Vertrauensprozessen zuschreiben zu können; siehe z. B. Kosfeld et al. (2005).

vertrauenswürdiger erachtet werden als Männer. So berichtete unter anderem die *New York Times*, dass in Mexiko-Stadt Teams weiblicher Verkehrspolizisten gebildet wurden, weil man glaubte, dass weibliche Beamte weniger wahrscheinlich Bestechungsgelder annehmen würden als männliche Beamte. Auf die Frage nach dem Grund für diese unkonventionelle Regelung antwortete der Polizeichef: „Ich vertraue ihnen“; diese Anekdote untermauert die Annahme, dass Frauen vertrauenswürdiger sind als Männer (Treaster 1999). Wissenschaftliche Befunde bestätigen sowohl die Annahme, dass Frauen weniger wahrscheinlich Bestechungsgelder annehmen, als auch die Annahme, dass Korruption mit dem Anteil der Frauen in der öffentlichen Verwaltung negativ korreliert ist (z. B. Dollar et al. 2001; Swamy et al. 2001).

Doch sind Frauen in Bezug auf ihr Verhalten tatsächlich vertrauenswürdiger als Männer? Obwohl es in der Vergangenheit in der wissenschaftlichen Gemeinschaft immer wieder viel beachtete Veröffentlichungen auf der Basis des Vertrauensspiels gab, die zeigen, dass Frauen vertrauenswürdiger sind als Männer (also in der Rolle des Vertrauensnehmers durchschnittlich höhere Beträge retournieren) (z. B. Buchan et al. 2008), so zeigt die Meta-Analyse von Van den Akker et al. (2020) *keine* geschlechterspezifischen Unterschiede in Bezug auf die Vertrauenswürdigkeit. Dieser Befund deckt sich mit den Aussagen in einem aktuellen narrativen Review (Sijtsma & Krabbendam 2022).

In Publikation Nr. 1 der vorliegenden Dissertation wird untersucht, ob es zwischen Frauen und Männern Hirnaktivierungsunterschiede bei der Verarbeitung unterschiedlich vertrauenswürdiger eBay-Websites gibt. Somit fokussiert diese Studie auf computerbasierte menschliche Interaktion (Verkäufer und Käufer interagieren über das Internet, Szenario B in Tabelle 2). Es ist daher in diesem Zusammenhang auch wichtig, auf Forschungsbefunde zu geschlechterspezifischen Unterschieden im Online-Shopping mit Bezug auf Vertrauen einzugehen (eine systematische Analyse dazu findet sich in Publikation Nr. 1).

Im Gegensatz zu Männern nehmen Frauen ein höheres Risiko beim Online Shopping wahr. Wenn eine Website von einem Freund empfohlen wird, führt dies bei Frauen zu einer stärkeren Verringerung des wahrgenommenen Risikos und einer stärkeren Bereitschaft, online einzukaufen, als bei Männern (Garbarino & Strahilevitz 2004). Zudem zeigt die Forschung, dass Frauen in Bezug auf Datenschutz beim Online Shopping ein höheres Risiko wahrnehmen als Männer (Sheehan 1999). Weiter wird darüber berichtet, dass der Effekt von Vertrauen auf die Absicht, online einzukaufen, bei Frauen stärker ausgeprägt ist als bei Männern (Awad & Ragowsky 2008). In einer Gesamtschau sind diese auf die digitale Welt fokussierten Befunde im Einklang mit dem Ergebnis, dass Männer grundsätzlich mehr vertrauen als Frauen. In einigen wenigen Studien wird zudem über geschlechterspezifische Hirnaktivierungsunterschiede beim Vertrauensspiel berichtet (siehe dazu eine aktuelle Synopse von

Sijtsma & Krabbendam 2022, S. 284-286). Die wesentliche Erkenntnis ist hierbei, dass sich diese Unterschiede auf die im Modell von Krueger und Meyer-Lindenberg (2019) dargestellten Areale beziehen (siehe Abbildung 5).

2.4.2 Vertrauen und persönlichkeitsspezifische Unterschiede

Die *American Psychological Association (APA)* definiert Persönlichkeit als

> *„the enduring configuration of characteristics and behavior that comprises an individual's unique adjustment to life, including major traits, interests, drives, values, self-concept, abilities, and emotional patterns. Personality is generally viewed as a complex, dynamic integration or totality shaped by many forces, including hereditary and constitutional tendencies; physical maturation; early training; identification with significant individuals and groups; culturally conditioned values and roles; and critical experiences and relationships" (APA 2023).*

Es ist offensichtlich, dass die Persönlichkeit eines Menschen mit seinen Vertrauenswahrnehmungen und -entscheidungen in Zusammenhang steht, weil sie beeinflusst, wie Menschen auf alle möglichen Stimuli reagieren (kognitions-, emotions- und verhaltensmäßig; Funder 2001) und somit auch auf solche, die in Vertrauenssituationen bedeutsam sind (z. B. Eigenschaften eines Vertrauensnehmers).

In der Fachliteratur wird die Meinung vertreten, dass das Big-Five-Modell (John et al. 2008; McCrae & Costa 1997; McCrae & Costa 1999) „the broadest level of abstraction" zur Konzeptualisierung von Persönlichkeit darstellt und sehr wahrscheinlich auch weltweit die höchste Bekanntheit hat (John & Srivastava 1999, S. 105). Jede der fünf Dimensionen (Offenheit für Erfahrungen, Gewissenhaftigkeit, Extraversion, Verträglichkeit, Neurotizismus) subsummiert spezifischere Persönlichkeitsmerkmale. Neben dem Big-Five-Modell (Revised NEO Personality Inventory (NEO PI-R) (McCrae et al. 2005) gibt es noch eine Reihe weiterer Konzeptualisierungen von Persönlichkeit, die in der Geschichte der Persönlichkeitspsychologie als bedeutsam beschrieben werden, einige wichtige davon sind: Myers-Briggs Type Indicator (MBTI) (Myers et al. 1998), Eysenck's Persönlichkeitsmodell (Eysenck 1947; Eysenck & Eysenck 1976), HEXACO-Modell (Ashton et al. 2004) sowie die revidierte Reinforcement Sensitivity Theory (rRST; Gray & McNaughton 2000).

Doch wie hängen Persönlichkeit und Vertrauen zusammen? Bevor auf diese Frage eingegangen wird, ist die Feststellung wichtig, dass die Vertrauensneigung („propensity to trust") als ein spezifisches Persönlichkeitsmerkmal konzeptualisiert ist, das dem abstrakteren Faktor *Verträglichkeit* aus dem Big-

Five-Modell als Facette zugeordnet ist (z. B. John & Srivastava 1999; John et al. 2008; McCrae & Costa 1997; McCrae & Costa 1999).

Freitag und Bauer (2016) haben den Zusammenhang von Persönlichkeit (Big-Five) und Vertrauen empirisch untersucht. Ihre Befunde zeigen: (1) der Einfluss der Persönlichkeit auf Vertrauen in Unbekannte ist größer als auf Vertrauen in Freunde, (2) Gewissenhaftigkeit und Offenheit sind bedeutsam für die Entwicklung von Vertrauen in Unbekannte *und* Freunde sowie (3) Verträglichkeit ist bedeutsam für die Entwicklung von Vertrauen in Unbekannte. Aus diesen Ergebnissen folgt, dass der Einfluss der Persönlichkeit auf das Vertrauen entscheidend davon abhängt, *wem* vertraut werden soll.

In Bezug auf zwischenmenschliches Vertrauen gibt es nach Freitag und Bauer (2016) zwei theoretische Perspektiven darüber, wie Vertrauen entsteht. Die *erste Perspektive* betont, dass das Vertrauen einer Person im Wesentlichen eine Funktion ihrer sozialen Umgebung ist. Daraus folgt, dass das situationsspezifische Vertrauen insbesondere auf den Erfahrungen und Lernprozessen in Bezug auf die Vertrauenswürdigkeit von Interaktionspartnern beruht (Coleman 1990; Hardin 2002). Basierend auf Erfahrungen prognostiziert somit ein Individuum das zu erwartende Verhalten anderer Personen. Freitag und Bauer (2016) geben jedoch an, dass sich Vertrauen auf gesellschaftlicher und individueller Ebene oft nur langsam verändert, was die Bedeutsamkeit späterer Lebenserfahrungen sowie situativer Anpassungen des Vertrauensverhaltens in Frage stellt. In Anbetracht dieses Umstands besagt die *zweite Perspektive*, dass Vertrauen eher eine stabile Neigung und somit ein situativ wenig veränderbares Persönlichkeitsmerkmal ist (Sztompka 1998; Uslaner 2018). Gemäß dieser zweiten Perspektive bewirken insbesondere individuelle Unterschiede von in der Kindheit entwickelten Persönlichkeitsmerkmalen Unterschiede im späteren Vertrauensverhalten. Zudem steht diese zweite Perspektive mit dem Einfluss genetischer Faktoren auf Vertrauenseinstellungen und -verhalten im Einklang, auch wenn Lernprozesse in der Kindheit einen maßgeblichen Einfluss auf das spätere Vertrauensverhalten haben.

Eine Sichtung der Fachliteratur zeigt, dass der Großteil der Forschung auf die erste Perspektive fokussiert und somit situative Erfahrungen und Lernprozesse als lebenslang vertrauensdeterminierend auffasst (das bedeutet, neue Erfahrungen bewirken immer wieder ein Update kognitiver Schemata; Bellucci & Dreher 2022). Die zweite Perspektive wurde seltener untersucht (Freitag und Bauer 2016), dennoch gibt es ein paar bedeutsame Forschungsbefunde. Mondak und Halperin (2008) berichten über einen positiven und statistisch signifikanten Zusammenhang zwischen Verträglichkeit und zwischenmenschlichem Vertrauen, ein statistisch signifikanter Zusammenhang mit anderen Persönlichkeitsmerkmalen des Big-Five-Modells konnte nicht gefunden werden. Befunde von Dinesen et al. (2014) zeigen, dass alle Big-Five-Merkmale einen statistisch signifikanten Zusammenhang mit generalisiertem Vertrauen aufweisen (Offenheit für Erfahrungen: positiv, Gewissenhaftigkeit: negativ, Extraversion: positiv, Verträglichkeit: positiv, Neurotizismus: negativ). Weiter zeigen die Ergebnisse von Hiraishi et al. (2008), dass alle Big-Five-Merkmale einen statistisch signifikanten Zusammenhang mit generalisiertem Vertrauen aufweisen (Offenheit für Erfahrungen: positiv, Gewissenhaftigkeit: positiv, Extraversion: positiv, Verträglichkeit: positiv, Neurotizismus: negativ). Pfattheicher und Böhm (2018) haben das Merkmal Ehrlichkeit-Bescheidenheit aus dem HEXACO-Modell und seinen Zusammenhang mit Vertrauen untersucht. Die Untersuchung fokussierte dabei darauf, zu zeigen, wie Selbstunsicherheit diesen Zusammenhang beeinflusst. Das Kernergebnis der Studie ist, dass Selbstunsicherheit den positiven Effekt von Ehrlichkeit-Bescheidenheit auf Vertrauen reduziert, und zwar deshalb, weil Selbstunsicherheit zu einer Reduktion positiver sozialer Erwartungen führt. Freitag und Bauer (2016) führen noch weitere empirische Forschungsergebnisse an, die zeigen, dass generalisiertes Vertrauen mit weiteren Persönlichkeitsmerkmalen in Beziehung steht, unter anderem mit Optimismus (positiver Zusammenhang) sowie mit Schüchternheit und Eifersucht (negativer Zusammenhang).

In einer Gesamtschau der hier vorgestellten Befunde zeigt sich, dass der Zusammenhang zwischen Persönlichkeit und Vertrauen komplex ist und mit vielen Kontextfaktoren interagiert (z. B. Saef et al. 2019). Da die gegenständliche Dissertation insbesondere auch Vertrauen in Digitaltechnologien wie KI-basierte Systeme untersucht, wurde ob der in diesem Bereich vergleichsweise spärlichen Befundlandschaft ein Review angefertigt (siehe Publikation Nr. 6), um folgende Frage zu beantworten: *Gibt es zwischen Vertrauen in KI-basierte Systeme und der Benutzerpersönlichkeit einen Zusammenhang?* Weiter wurde eine empirische Studie im Kontext der Nutzung KI-basierter Systeme in der Medizin (Arzt-Patienten-Interaktion) durchgeführt. Es wurden hierbei auch mehrere Merkmale zu individuellen Unterschieden erfasst (Geschlecht, Persönlichkeit, Alter, Bildungsstand, Vertrauensneigungen in Bezug auf Menschen sowie in Bezug auf Maschinen, Einstellung zu Digitaltechnologien, KI-Angst) und es wurde ermittelt, ob diese einen Zusammenhang mit Vertrauen und Misstrauen in Diagnosen und Therapievorschläge (und mit weiteren Variablen) in Abhängigkeit davon aufweisen, ob der Arzt ein Mensch oder ein Chatbot ist (zudem wurde die Zwischenstufe menschlicher Arzt, der durch ein KI-System unterstützt wird, untersucht) (siehe dazu Publikation Nr. 7).

2.4.3 Vertrauen und Einflüsse durch die Parkinson-Krankheit

In der Fachliteratur wird darüber berichtet, dass zwischen Vertrauen und diversen Neuropathologien ein Zusammenhang besteht. Drei Beiträge in einem aktuellen Herausgeberband zur Neurobiologie des Vertrauens (Krueger 2022) fassen wissenschaftliche Befunde in folgenden drei Bereichen synoptisch zusammen: Vertrauen und Psychose (Lemmers-Jansen & Fett 2022), Vertrauen und

Persönlichkeitsstörungen (Lis et al. 2022) sowie Vertrauen und Hirnläsionen (Wadsworth & Tranel 2022).

Bemerkenswert ist, dass die Parkinson-Krankheit bis zu jener im Rahmen der vorliegenden Dissertationsschrift vorgestellten Untersuchung (siehe Publikation Nr. 4) nicht mit Vertrauensverhalten in Verbindung gebracht wurde. Wie bereits in Kapitel 2.3 bei den neurobiologischen Grundlagen von Vertrauen ausgeführt, zeigen Forschungsbefunde, dass Vertrauenswahrnehmungen und -verhalten mit Aktivierung in spezifischen Hirnregionen einhergehen (siehe dazu die Reviews von Krueger & Meyer-Lindenberg 2019 und Riedl & Javor 2012), und es sind überwiegend genau diese Hirnregionen, die durch die Parkinson-Krankheit in ihrer Struktur und Funktion pathologisch verändert sind (Javor et al. 2015), konkret die Basalganglien (die unter anderem für Belohnungs- und Lernprozesse zentral sind), das limbische System (das unter anderem für Risiko- und Unsicherheitswahrnehmungen zentral ist) und der Frontalkortex (der unter anderem für Mentalizing-Prozesse, auch als Theory-of-Mind-Prozesse bekannt; Baron-Cohen 1999) zentral ist; Mentalizing bezeichnet die Fähigkeit, die Gedanken und Absichten anderer Menschen zu verstehen).

Vertrauen ist insbesondere auch in der Arzt-Patienten-Beziehung bedeutsam, da die Befolgung von Therapiervorschlägen durch den Patienten wie die Einnahme bestimmter Medikamente durch das Vertrauen in den Arzt beeinflusst wird (z. B. Grosset et al. 2005; Thom & Campbell 1997). Nicht zuletzt aus diesem Umstand ergab sich die Motivation zur Durchführung der in Publikation Nr. 4 vorgestellten Untersuchung. Wenn gezeigt werden kann, dass Menschen mit Parkinson-Erkrankung im Vergleich zu gesunden Kontrollprobanden weniger vertrauen (z. B. gemessen auf der Basis des Vertrauensspiels), dann wäre dies eine mögliche Erklärung für die – relativ zu anderen Krankheiten betrachtet – schwächer ausgeprägte Rate der Befolgung von Therapievorschlägen, insbesondere der Einnahme von Medikamenten zur Kompensation der pathologisch reduzierten zentralnervösen Dopaminspiegel wie L-DOPA (Javor et al. 2015).

Zu erwähnen ist in diesem Zusammenhang auch, dass bei der Durchführung eines solchen Laborexperiments explizit darauf zu achten ist, individuelle Unterschiede zwischen der Gruppe der Patienten und der Gruppe der gesunden Kontrollprobanden zu beachten, unter anderem Alter, Geschlecht, Bildungsstand sowie Einkommen. Zudem gilt es zu berücksichtigen, dass in der Gruppe der Patienten möglichst keine Personen sind, die unter neuropsychiatrischen Begleiterkrankungen leiden (z. B. kognitive und/oder affektive Störungen, was bei der Parkinson-Krankheit nicht einfach ist). Wäre dies der Fall, dann ist unklar, ob die Pathologie der Parkinson-Erkrankung selbst und somit die daraus folgenden Konsequenzen für die Struktur und Funktion vertrauensrelevanter Hirnareale oder die daraus gegebenenfalls resultierenden neuropsychiatrischen Begleiterkrankungen ursächlich für Vertrauensdefizite sind (möglich ist auch, dass beide Faktoren Varianz eines Vertrauensdefizits aufklären). Es ist bedeutsam, solche bei der Parkinson-Erkrankung relevanten Begleiterkrankungen zu messen (im Regelfall auf der Basis von Fragebögen), um Menschen mit pathologischen Werten auszuschließen (Javor et al. 2015). Ein Beispielfragebogen ist der QUIP (Questionnaire for Impulsive-Compulsive Disorders) (Weintraub et al. 2009).

3. Publikationen und Erkenntnisse

Im Folgenden werden die sieben dieser Dissertationsschrift zugrunde liegenden Publikationen zusammengefasst. Die Originalpublikationen sind in Kapitel 6 zu finden.

3.1 Publikation 1 (Riedl et al. 2010), Laborexperiment: Gibt es neuronale Geschlechterunterschiede im Online-Vertrauen? Eine fMRT-Studie zur wahrgenommenen Vertrauenswürdigkeit von eBay-Angeboten

In Kapitel 2.4.1 wurden bereits wichtige Erkenntnisse über Vertrauen und geschlechterspezifische Unterschiede zusammengefasst, so auch zu Einstellungs- und Verhaltensunterschieden im Bereich Online Shopping. Zum Zeitpunkt der Erarbeitung von Publikation Nr. 1 erklärte jedoch die Forschung im Bereich der Informationssysteme nicht zufriedenstellend, *warum* diese geschlechtsspezifischen Unterschiede existieren. Als möglicher Grund wurde angeführt, dass Forschungsansätze oft nicht den offensichtlichsten Faktor berücksichtigen, der menschliches Verhalten beeinflusst: die Biologie. Angesichts der Bedeutsamkeit von neurobiologischen Faktoren bei Entscheidungen über Vertrauen sollten die biologischen Einflüsse natürlich auch jene in Bezug auf das Geschlecht einschließen. Um empirisch zu zeigen, dass Vertrauen im Online Shopping mit Veränderungen der Aktivität in bestimmten Gehirnregionen in Verbindung steht, die sich zwischen Männern und Frauen signifikant unterscheiden, wurde eine fMRT-Studie durchgeführt. In einem Laborexperiment wurde die Gehirnaktivität von 10 weiblichen und 10 männlichen gesunden Probanden während dem Fällen von Vertrauensentscheidungen zu eBay-Angeboten erfasst.

Eine wesentliche Erkenntnis der Studie war, dass bei Frauen und Männern größtenteils unterschiedliche Hirnregionen bei den Vertrauensentscheidungen aktiv waren. Diese Ergebnisse stehen im Einklang mit der *Empathizing-Systemizing-Theorie* nach Baron-Cohen et al. (2005), die geschlechtsspezifische Unterschiede im neuronalen Informationsverarbeitungsmodus vorhersagt. Ursprünglich wurde diese Theorie im Bereich der Autismusforschung entwickelt. Die Theorie besagt unter anderem, dass Autismus ein Extrem des männlichen Informationsverarbeitungsmodus darstellt; konkret – in Relation zu Frauen – reduzierte Empathie und gesteigerte Systematisierung. Empathizing ist die Fähigkeit, das Verhalten von anderen Menschen vorherzusagen und darauf zu reagieren, indem man die mentalen Zustände anderer Akteure ableitet und mit angemessenen Emotionen reagiert.

Systemizing ist die Fähigkeit, das Verhalten deterministischer Systeme vorherzusagen und darauf zu reagieren, indem man Input-Output-Beziehungen analysiert und die Regeln ableitet, die solche Systeme steuern. Die Empathizing-Systemizing-Theorie sagt voraus, dass Frauen – auf Bevölkerungsebene – stärker im Empathizing sind und Männer stärker im Systemizing.

Ein weiteres Ergebnis ist, dass bei Frauen mehr Gehirnregionen aktiviert waren als bei Männern. Dieses Ergebnis steht im Einklang mit dem *Selektivitätsmodell* nach Meyers-Levy und Kollegen (Meyers-Levy 1994; Meyers-Levy & Maheswaran 1991; Meyers-Levy & Sternthal 1991). Dieses Modell besagt, dass Männer – vergleichsweise zu Frauen – öfter *keine* umfassende Verarbeitung der verfügbaren Informationen durchführen, sondern eher selektiv vorgehen; das heißt, sie betrachten und verlassen sich nur auf Teilmengen der zur Verfügung stehenden Informationen und aktivieren dabei weniger Gehirnbereiche im Vergleich zu Frauen. Frauen neigen hingegen dazu, umfassendere Informationsverarbeitungsstrategien zu verwenden, indem sie eine größere Menge verfügbarer Informationen berücksichtigen und eine aufwändigere und differenziertere Analyse aller verfügbaren Informationen durchführen. Im Einklang mit diesem Argument fand eine Studie (Rodgers und Harris 2003) im Bereich des Online-Shoppings heraus, dass Männer im Gegensatz zu Frauen eine Einkaufsseite typischerweise anhand einer allgemeinen Einstellung beurteilen. Frauen hingegen achten in der Regel auf die detaillierteren Aspekte einer Website und stützen ihr Urteil mehr darauf, wie diese Informationen sie emotional beeinflussen. Dies geht offensichtlich mit der Aktivierung einer größeren Anzahl von Gehirnbereichen einher. Die Aussagen des Selektivitätsmodells sind auch mit der höheren Neigung von Frauen zur Risikowahrnehmung kompatibel. Wie bereits in Kapitel 2.4.1 beschrieben nehmen Frauen beim Online-Kauf im Vergleich zu Männern in der Regel ein höheres Maß an Risiko wahr (z. B. Garbarino und Strahilevitz 2004). Eine mögliche Strategie für Konsumentinnen, mit dieser höheren Risikowahrnehmung umzugehen, besteht darin, nach mehr Informationen zu suchen und diese zu berücksichtigen, da ein Mehr an Information Unsicherheit reduzieren kann (Hall et al. 2007).

Am fMRT-Experiment nahmen Probanden im Alter von 30 bis 35 Jahren teil und es wurden etliche Kontrollvariablen erfasst, um auszuschließen, dass diese Faktoren die Vertrauensergebnisse beeinflussen (z. B. Erfahrung mit dem Internet, Einstellung zu eBay, generalisiertes Vertrauen; Details dazu finden sich in Publikation Nr. 1, S. 408-410). Die Aufgabe für die Probanden im Kernspintomographen bestand darin, insgesamt 120 eBay-Angebote (ein Beispiel ist in Abbildung 6 dargestellt) für jeweils 12 Sekunden zu betrachten, um danach auf der Basis einer MRT-kompatiblen Antwortbox anzugeben, ob sie das jeweilige Angebot für vertrauenswürdig oder nicht vertrauenswürdig hielten. Nach dem 12-Sekunden-Angebotsbild sahen die Teilnehmer für 3 Sekunden ein Fixationskreuz. Dann wurde das nächste Angebot präsentiert und die Anzeigen setzten sich in dieser Weise fort. Die Reihenfolge der Angebote war randomisiert.

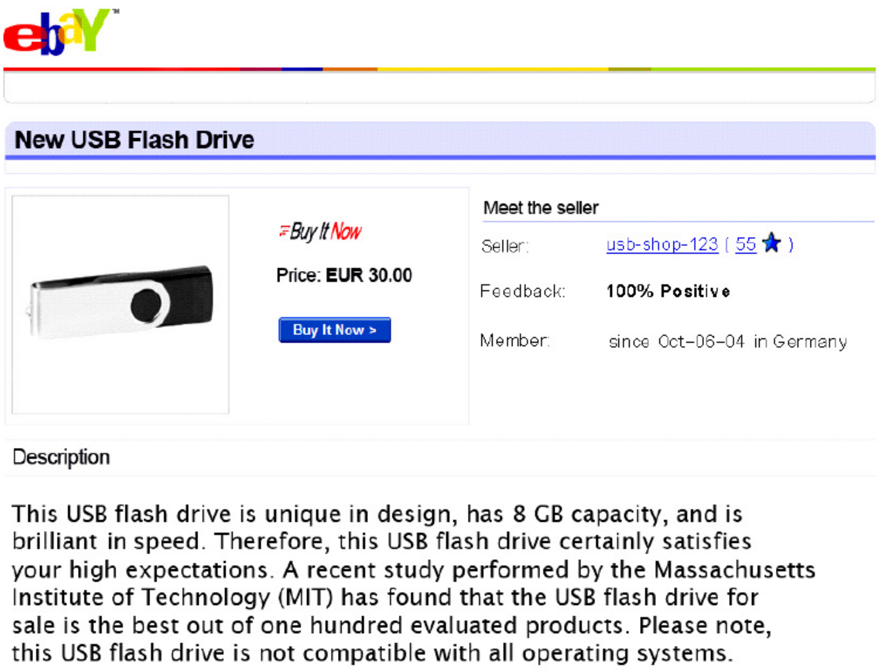


Abbildung 6: Beispiel-Stimulus der fMRT-Studie, der ein vertrauenswürdiges eBay-Angebot zeigt (Quelle: Riedl et al. 2010, S. 408)

Tabelle 3 fasst wesentliche Ergebnisse der Studie zusammen (diese basieren auf Whole-Brain-Analysen). In der linken Spalte sind die bedeutsamsten Hirnregionen im Vertrauenskontext angeführt und in der rechten Spalte die geschlechterspezifischen Ergebnisse der Studie (in den weiteren Spalten sind die Ergebnisse wichtiger verwandter Arbeiten dargestellt). Die folgenden Hirnregionen waren in der vorliegenden Studie nur bei Frauen aktiviert: Ventral Anterior Cingulate Cortex (BA 24), Caudate Nucleus, Hippocampus und Parahippocampus Gyrus, Putamen und der Thalamus. Die folgenden Hirnregionen waren nur bei Männern aktiviert: Ventral Posterior Cigulate Cortex (BA 23) und Ventromedial Prefrontal Cortex (BA 10). Die weiteren benannten Regionen waren bei Frauen und Männern aktiv: Dorsolateral Anterior Cingulate Cortex (BA 32), Dorsolateral Prefrontal Cortex (BA 9) und Insular Cortex. Eine Aktivierung der Amygdala wurde in der gegenständlichen Studie nicht gefunden.

	Baumgartner et al. (2008) (Trust Game)	Delgado et al. (2005) (Trust Game)	King-Casas et al. (2005) (Trust Game)	Krueger et al. (2007) (Trust Game)	Winston et al. (2002) (Trustworthiness of Human Faces)	Present Study (Trustworthiness of Internet Offers)
Sample Size (Female/Male)	**49 (0/49)**	**12 (5/7)**	**96 (n.a.)**	**44 (22/22)**	**14 (6/8)**	**20 (10/10)**
Amygdala	•				•	
Ventral Anterior Cingulate Cortex (BA 24)	•		•			F
Caudate Nucleus	•	•	•			F
Dorsolateral Anterior Cingulate Cortex (BA 32)		•		•		F/M
Dorsolateral Prefrontal Cortex (BA 9)				•		F/M
Hippocampus and Parahippocampus Gyrus		•			•	F
Insular Cortex	•	•			•	F/M
Putamen	•	•				F
Thalamus	•		•		•	F
Ventral Posterior Cingulate Cortex (BA 23)						M
Ventromedial Prefrontal Cortex (BA 10)						M

Notes: The table lists brain regions that are mentioned in at least two of the six studies. Exceptions are the ventral posterior cingulate cortex (BA 23) and the ventromedial prefrontal cortex (BA 10), which were identified in the present study. Only the present study reports gender-specific results. BA: Brodmann Area. F: Female. M: Male. n.a.: not available.

Tabelle 3: Ergebnisse der fMRT-Studie und Einordnung in den zum Zeitpunkt der Publikation existierenden Forschungsstand (Quelle: Riedl et al. 2010, S. 414)

Indem gezeigt werden konnte, dass die wahrgenommene Vertrauenswürdigkeit von Internetangeboten durch Hirnaktivierungsmuster, die sich zwischen Frauen und Männern signifikant unterschieden haben, beeinflusst wird, hat diese Studie bedeutende Auswirkungen auf die Forschung im Bereich Informationssysteme (IS).[12] Es wurde der Erkenntniswert eines Forschungsansatzes bestätigt (konkret: neuronale Korrelate von Entscheidungen sichtbar zu machen), der zum damaligen Zeitpunkt in der IS-Forschung vernachlässigt wurde. Mittlerweile hat dieser als NeuroIS bekannte Forschungsansatz massiv an Bedeutung gewonnen (siehe www.NeuroIS.org).

3.2 Publikation 2 (Riedl & Javor 2012), Review: Die Biologie des Vertrauens: Integration von Forschungsbefunden aus Genetik, Endokrinologie und funktioneller Hirnbildgebung

In Kapitel 2.3 wurden bereits wichtige neurobiologische Grundlagen von Vertrauen dargestellt. Insbesondere wurde auf ein Modell zum zwischenmenschlichen Vertrauen („Model of Interpersonal Trust"), das auf der Basis von Erkenntnissen aus den Neurowissenschaften, der Psychologie und der Ökonomie entwickelt wurde, eingegangen. Zum Zeitpunkt der Erarbeitung von Publikation Nr. 2 existierte dieses Modell zum zwischenmenschlichen Vertrauen von Krueger und Meyer-Lindenberg (2019) jedoch noch nicht. Vielmehr referenzieren Krueger und Meyer-Lindenberg das von Riedl und

[12] Die Wirtschaftsinformatik ist im deutschsprachigen Raum jene Wissenschaftsdisziplin, die als Pendant zur internationalen Information Systems Discipline angesehen wird (Heinzl et al. 2024).

Javor (2012) entwickelte Modell, das Teil dieser Dissertationsschrift ist. Im Folgenden werden die wesentlichen Erkenntnisse von Publikation Nr. 2 dargelegt.

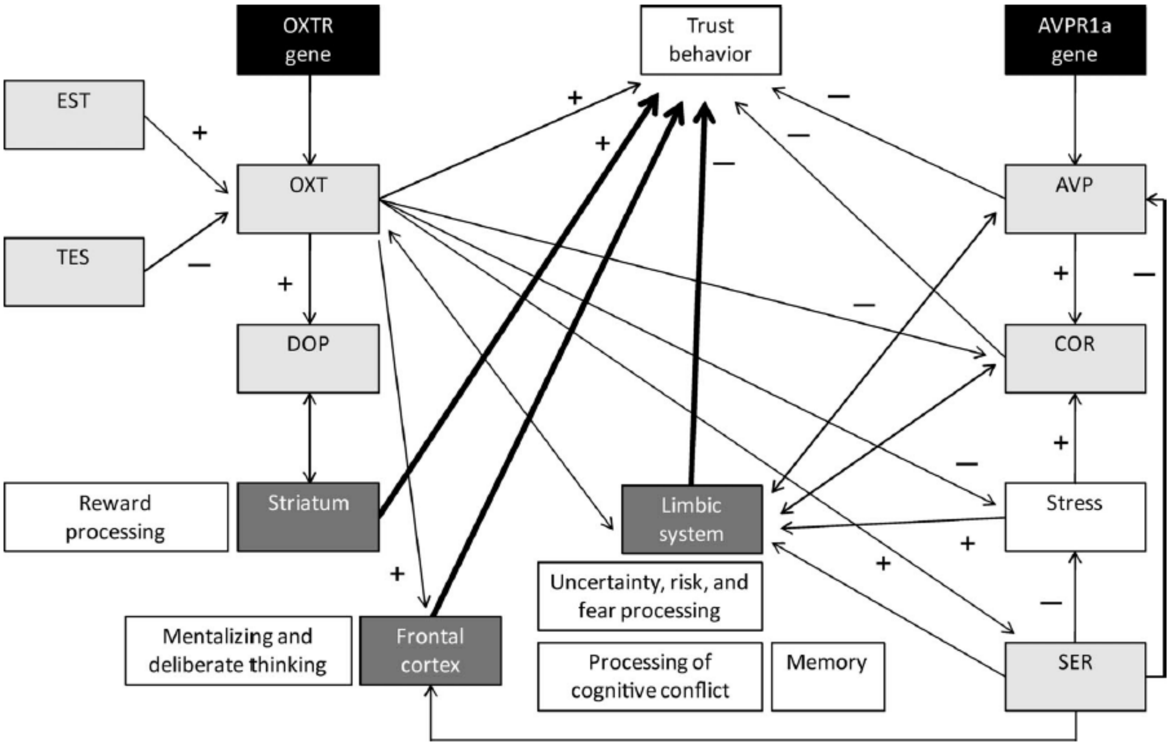


Figure 5. Conceptual Model of the Biology of Trust. *Notes:* AVP: Arginine Vasopressin, AVPR1a: Arginine Vasopressin 1a Receptor, COR: Cortisol, DOP: Dopamine, EST: Estrogen, OXT: Oxytocin, OXTR: Oxytocin Receptor, SER: Serotonin, TES: Testosterone. Black boxes indicate a genetic level of analysis. Light gray boxes indicate a hormonal level of analysis. Dark gray boxes indicate a brain level of analysis. White boxes indicate a behavioral level of analysis (trust behavior and stress), as well as cognitive constructs (reward processing; uncertainty, risk, and fear processing; processing of cognitive conflict; mentalizing and deliberate thinking). Plus (+) and minus (−) signs indicate the hypothesized valence of the relation (positive or negative). Direction of the arrows indicates the hypothesized causality. A two sided-arrow indicates interplay among two variables. The arrows in bold illustrate that the influences of genes and hormones on trust behavior are mediated by activation in specific brain areas. Thus, a direct link between a hormone and trust behavior indicates that research has investigated this relation without considering brain activity as a mediator (e.g., OXT→Trust behavior, Kosfeld et al., 2005). Anatomical connections between the brain areas (striatum, limbic system, and frontal cortex) are not illustrated to facilitate clarity of the model.

Abbildung 7: Konzeptionelles Modell zur Biologie des Vertrauens (Quelle: Riedl und Javor 2012, S. 80)

Abbildung 7 zeigt das entwickelte Modell zur Biologie des Vertrauens. Das Modell konzeptualisiert drei Ebenen: Gehirn (dunkelgraue Rechtecke), Hormone (hellgraue Rechtecke), Gene (schwarze Rechtecke) und psychologische Konzepte (weiße Rechtecke). Ein wesentliches Ziel der Modellentwicklung war es, die Beziehungen zwischen den in Vertrauenssituationen relevanten psychologischen Konzepten und den drei biologischen Ebenen transparent zu machen. Zudem werden Beziehungen zwischen den drei biologischen Faktoren dargestellt. Die Entwicklung des Modells erfolgte auf der Basis einer Analyse einer dreistelligen Anzahl von Forschungsbefunden. Die wesentlichen Erkenntnisse dieser Review-Arbeit werden im Folgenden zusammengefasst (die den Ausführungen zugrunde liegenden Originalstudien finden sich in Riedl und Javor 2012, siehe insbesondere S. 78-82, sowie in der dortigen Referenzliste).

Jeder Mensch hat eine bestimmte genetische Veranlagung, die auch das Oxytocin-Rezeptor (OXTR)-Gen und das Arginin-Vasopressin-1a-Rezeptor (AVPR1a)-Gen betrifft. Die Eigenschaften dieser Gene beeinflussen direkt die Signaltransduktion zweier Hormone, nämlich Oxytocin (OXT) und Arginin-Vasopressing (AVP). Während hohe OXT-Spiegel tendenziell zu vertrauensvollem Verhalten führen, resultieren hohe AVP-Spiegel typischerweise in misstrauischem Verhalten – OXT und AVP sind Antagonisten (Donaldson & Young 2008). Wie im Modell veranschaulicht, wird OXT von Sexualhormonen beeinflusst. Erstens stehen EST und OXT in einem positiven Zusammenhang (Verbalis 1999). Zweitens weist das männliche Sexualhormon Testosteron (TES) antagonistische Eigenschaften gegenüber OXT auf. Daher wirkt sich TES negativ auf OXT aus (Bos et al. 2010).

Weiter wurde gezeigt, dass OXT den Dopamin(DOP)-Spiegel erhöht, der eine entscheidende Rolle bei der Belohnungsverarbeitung spielt. Diese positive Beziehung zwischen den beiden Hormonen geht mit einer entsprechenden Beziehung auf Kognitions- und Verhaltensebene einher, da das Ziel, jemandem zu vertrauen, in engem Zusammenhang mit der Erwartung einer Belohnung steht. Dopamin ist somit fundamental mit der Belohnungsverarbeitung im Gehirn korreliert. In Übereinstimmung damit fanden eine Reihe von fMRT-Experimenten in Vertrauenssituationen eine Aktivierung im ventralen Striatum, insbesondere im Nucleus Accumbens. Daher ist eine wesentliche Aussage des Modells, dass eine Aktivierung im ventralen Striatum in der Regel zu vertrauensvollem Verhalten führt.[13]

Abbildung 7 veranschaulicht weiter, dass OXT die Fähigkeit positiv beeinflusst, die inneren Zustände anderer Menschen abzuleiten (beispielsweise ihre Absichten, Gedanken und Gefühle), um ihr Verhalten vorherzusagen. Diesem Mentalizing-Prozess liegt insbesondere Aktivierung im Frontalkortex zugrunde, insbesondere im medialen Frontalkortex. Zu beachten ist hierbei, dass das Mentalizing ein grundlegender kognitiver Prozess in Vertrauenssituationen ist, da die Entscheidung zu vertrauen impliziert, über die Absichten eines Interaktionspartners nachzudenken, um dessen Vertrauenswürdigkeit besser einzuschätzen. Daher beeinflusst das Mentalizing und die entsprechende Aktivität im Frontalkortex das Vertrauensverhalten.

OXT steht auch in Zusammenhang mit dem limbischen System. Limbische Regionen, die mit Emotionen und Gedächtnis in Verbindung stehen (z. B. Amygdala, Hippocampus und Hypothalamus), haben eine Ansammlung von OXT-Rezeptoren (z. B. Heinrichs et al. 2004, Loup et al. 1991). Zudem

[13] Die Forschung zeigt, dass auch das mesokortikale Dopaminsystem mit der Belohnungsverarbeitung zusammenhängt (z. B. Hauser et al. 2017, Tzschentke 2001).

ist bekannt, dass OXT auf limbische Regionen, insbesondere auf die Amygdala, wirkt, indem es erregende Informationen von diesen Regionen zu Hirnstammstellen hemmt, die autonome Angstreaktionen vermitteln (Debiec 2005, Huber et al. 2005, Petrovic et al. 2008). Neben den Hirnstrukturen, die mit der Verarbeitung negativer Emotionen wie Angst (z. B. Amygdala) sowie Unsicherheit und Risiko (z. B. Insular Cortex) in Verbindung stehen, wurde gezeigt, dass eine bestimmte limbische Region die Verarbeitung kognitiver Konflikte neuronal implementiert, nämlich der Anterior Cingulte Cortex (ACC) (Botvinick et al. 1999, 2004). Da emotionsbezogene Hirnstrukturen wie die Amygdala reichlich Verbindungen zum ACC aufweisen, kann angenommen werden, dass die automatische und in der Regel unbewusste Informationsverarbeitung, insbesondere in der Amygdala, von der Verarbeitung kognitiver Konflikte im ACC begleitet wird, als Ergebnis des Ausgleichs zwischen den schnellen viszeralen Wahrnehmungen und bewussteren Gedanken (Riedl et al. 2010). Weiter zeigt die Analyse im Zuge der Review-Arbeit, dass die Amygdala und der Hippocampus bei Vertrauensentscheidungen wichtig sind. Beide limbischen Strukturen sind essentiell für die Wahrnehmung und Verarbeitung von Emotionen. Zusätzlich zu dieser allgemeinen Rolle in der Emotionswahrnehmung und -verarbeitung wurde sowohl die Amygdala als auch der Hippocampus spezifisch mit emotionsbasiertem Gedächtnis in Verbindung gebracht (Gazzaniga et al. 2009), was insbesondere beim Erleben eines Vertrauensmissbrauchs von Bedeutung ist (z. B. Adolphs et al. 1998, Baumgartner et al. 2008).

Abbildung 7 zeigt auch ein Zusammenspiel zwischen dem limbischen System und Kortisol (COR). Es ist bekannt, dass sozialer Stress zu erhöhter Amygdala-HPA-Aktivierung führt, was wiederum COR ansteigen lässt (Takahashi et al. 2005). Darüber hinaus beeinflusst COR das limbische System, da es die Aktivität des Hippocampus beeinflusst. Insbesondere wenn Erregungs- und Stressniveaus zu hoch werden, was zu einer ausgeprägten Erhöhung von COR führt, ist es möglich, dass das (soziale) Gedächtnis beeinträchtigt wird (Takahashi et al. 2004). Hinsichtlich der Beziehung zwischen OXT und COR (Stress) wurde festgestellt, dass intranasal verabreichtes OXT Speichel-COR-Werte signifikant reduziert (Ditzen et al. 2009, Heinrichs et al. 2003). In Übereinstimmung mit diesem Befund finden sich in der Fachliteratur auch Hinweise auf eine negative Korrelation zwischen zwischenmenschlichem Vertrauen und COR (Takahashi et al. 2005). Personen mit einem höheren Grad an zwischenmenschlichem Vertrauen haben eine schwächer ausgeprägte Kortisolveränderung bei sozialem Stress (mit dem Trier Social Stress Test induziert) (Takahashi et al. 2005). Zusätzliche Unterstützung für diese Ergebnisse liefern Befunde, die zeigen, dass Personen mit sozialen Phobien und Angststörungen, die normalerweise eine überschießende HPA-Reaktion auf sozialen Stress zeigen, hohe Werte beim zwischenmenschlichen Misstrauen aufweisen (Furlan et al. 2001, Takahashi et al. 2005).

Es wurde bereits dargelegt, dass hohe AVP-Spiegel typischerweise zu misstrauischem Verhalten führen (Heinrichs & Domes 2008, Heinrichs et al. 2009). Eine mögliche Erklärung für diesen Effekt liefert das limbische System (Debiec 2005). Es wurde gezeigt, dass eine erhöhte AVP-Expression im hypothalamischen paraventrikulären Kern mit erhöhten Angstniveaus verbunden ist, die wiederum mit der Aktivität der Amygdala zusammenhängen (Murgatroyd et al. 2004). In Übereinstimmung mit diesem Befund wird berichtet, dass AVP (a) mit erhöhter Wachsamkeit, Angst, Erregung und Aktivierung korreliert (Heinrichs & Domes 2008) und (b) die Kodierung sozialer Informationen verstärkt, um diese merkbarer zu machen (Guastella et al. 2010). Da sowohl COR als auch AVP das Vertrauensverhalten negativ beeinflussen, überrascht es nicht, dass Forschungsergebnisse auch eine direkte positive Beziehung zwischen diesen beiden Hormonen gefunden haben. Konkret wurde gezeigt, dass die intranasale Verabreichung von AVP die Speichel-COR-Werte bei einem sozialen Stresstest erhöht (de Winter et al. 2003, Ebstein et al. 2009, Patchev & Almeida 1995).

Ein weiteres Hormon, das in Abbildung 7 thematisiert wird, ist Serotonin (5-Hydroxytryptamin; 5-hat; im Folgenden SER). Es ist bekannt, dass ein niedriger SER-Spiegel zu einer höheren Ablehnungsrate unfairer Angebote in ökonomischen Spielen führt, was auf die Bedeutung von SER für die Kontrolle negativer Emotionen hinweist, die sowohl im limbischen System als auch im Frontalkortex verarbeitet und kontrolliert werden (Crockett et al. 2008, Sanfey et al. 2003). Diese negativen Emotionen spielen eine wichtige Rolle in Vertrauenssituationen (Winston et al. 2002). In diesem Zusammenhang hat die Forschung gezeigt, dass Personen mit niedrigeren SER-Werten im Gehirn weniger kooperativ in Wirtschaftsspielen sind als Kontrollpersonen (Wood et al. 2006). Darüber hinaus ist es naheliegend, dass eine negative Korrelation zwischen SER und Stress besteht; konkret haben Forschungsergebnisse gezeigt, dass Berührungen mit erhöhten SER-Werten einhergehen und Berührungen können die Stressreaktivität bei Menschen reduzieren (Field et al. 1996, Hanley & Van de Kar 2003). Schließlich wird berichtet, dass SER den AVP-Rezeptor desensibilisiert, was zu einer Reduzierung von Annäherungs- und Vertrauensverhalten führen könnte (Skuse & Gallagher 2011).[14]

Zusammenfassend wird nochmals festgehalten, dass das in Abbildung 7 dargestellte Modell zum Zeitpunkt seiner Publikation im Jahr 2012 zu den reichhaltigsten Modellen seiner Art im Vertrauensbereich zählte und bis heute Einfluss auf weiterführende Arbeiten hat, wie beispielsweise auf das 2019 von Krueger und Meyer-Lindenberg veröffentlichte Modell zum zwischenmenschlichen Vertrauen („Model of Interpersonal Trust“).

[14] Serotonin steht zudem in einem inversen Zusammenhang mit Depression (z. B. Cowen & Browning 2015) und Aggression (z. B. Duke et al. 2013).

3.3 Publikation 3 (Riedl et al. 2014), Laborexperiment: Vertrauen in Menschen und Avatare: Eine Hirnbildgebungsstudie basierend auf der Evolutionstheorie

Avatare sind künstliche Repräsentationen von Menschen oder Maschinen in der virtuellen Welt. Beispielsweise repräsentieren sich Menschen in virtuellen Umgebungen mit Avataren (z. B. in Videospielen oder im Metaverse[15]) und Unternehmen sowie öffentliche Institutionen verwenden Avatare auf ihren Websites als virtuelle Berater (z. B. https://www.linz.at/db/chatbot/ – dieser Chatbot der Stadt Linz namens ELLI hat den Zweck, Bürgerinnen und Bürger zu beraten). Wie in Tabelle 2 (Szenario D) dargelegt, gewinnt in einer zunehmend digitalen Welt die computerbasierte Interaktion zwischen Mensch und Digitaltechnologie (Maschine) massiv an Bedeutung. Insbesondere interagieren heutzutage immer mehr Menschen mit KI-Systemen, die in der Rolle eines virtuellen Beraters bzw. Kommunikationspartners agieren.

Vor dem Hintergrund dieser technologischen Entwicklungen wurden drei Forschungsfragen formuliert: (1) Haben Menschen unterschiedliche Fähigkeiten in Bezug auf die Unterscheidung der Vertrauenswürdigkeit von Mensch und Avatar? (2) Falls ja, gehen diese Unterschiede in der Urteilsfähigkeit mit neuronalen Unterschieden einher? (3) Gibt es Unterschiede im Lernen von Vertrauenswürdigkeit in Abhängigkeit davon, ob mit einem Mensch oder Avatar interagiert wird? Um diese Fragen zu beantworten, wurde ein Laborexperiment auf der Basis von fMRT und dem Vertrauensspiel durchgeführt.

Als theoretischer Rahmen für die Studie fungierte die *Evolutionäre Psychologie*, deren Erkenntnispotenzial erst in der jüngeren Vergangenheit auch für die Informationssystem-Forschung erkannt wurde (Kock 2004, 2005, 2009). Die Evolutionstheorie postuliert unter anderem, dass die Spezies Mensch sich durch natürliche Selektion entwickelt hat, ein Prozess, der über Millionen von Jahren stattgefunden hat und bei dem zufällige Mutationen in der genetischen Ausstattung von Nachkommen auftreten. Solche Mutationen beeinflussen Eigenschaften, die die Überlebens- und Reproduktionschancen erhöhen oder verringern, und werden an die Nachkommen weitergegeben, bis die Mutationen als charakteristische Eigenschaft einer Spezies etabliert sind (Darwin 1859). Der außerordentlich lange Hals und die Beine der Giraffe sind beispielsweise das Ergebnis eines solchen evolutionären Prozesses, weil diese Eigenschaften die Nahrungsaufnahme in Savannenumgebungen erleichterten (man denke an das Erreichen von Blättern in der Baumkrone) und somit die Fitness dieser Spezies erhöhte (gemessen an der Anzahl überlebender Nachkommen). Das Wahrnehmen

[15] Laut dem *Gabler Wirtschaftslexikon* ist das Metaverse wie folgt definiert: „Das Metaverse (dt. Metaversum) ist ein virtueller Raum, in dem sich Benutzer mit Hilfe von Avataren bewegen und in dem sie virtuelle Artefakte beeinflussen und nutzen können, etwa wenn sie sich Kleidung überziehen, ein Haus bauen und dieses einrichten, eine Tür öffnen und auf die Straße hinaustreten und dort Mitspieler und Gleichgesinnte treffen. Wie in der realen Welt kann man dort leben, arbeiten, lernen, Handel treiben, Gespräche führen und Beziehungen aufbauen." (Quelle: https://wirtschaftslexikon.gabler.de/definition/metaverse-123520, Abruf am 25.03.2024).

menschlicher Gesichtsinformationen, um die Vertrauenswürdigkeit einer anderen Person vorherzusagen, ist gleichfalls eine Eigenschaft, die sich im Laufe der Zeit entwickelt hat und die Überlebenschancen in Umgebungen erhöht hat, in denen soziale Interaktion ausschließlich von Angesicht zu Angesicht stattfand. Da diese Eigenschaft im Allgemeinen bei allen gesunden Individuen unabhängig vom kulturellen Hintergrund zu finden ist, kann sie als universelle menschliche Eigenschaft betrachtet werden (Kock 2009).

Die Universalität der Fähigkeit, Gesichtsinformationen zur Vorhersage der Vertrauenswürdigkeit zu verwenden, kann jedoch nicht unmittelbar auf Avatare übertragen werden. Da diese „virtuellen Menschen" als technologisches Artefakt erst seit wenigen Jahrzehnten entwickelt werden und somit erst seit kurzem existieren, ist der Zeitraum viel zu kurz, um zu Änderungen der Allelfrequenzen im menschlichen Genom zu führen (Boaz & Almquist 2001; Cartwright 2000; Kock & Hantula 2005). Vor diesem Hintergrund wurde folgende Hypothese formuliert: Basierend auf Gesichtsinformationen sind Menschen besser in der Lage, die Vertrauenswürdigkeit von Menschen vorherzusagen als die Vertrauenswürdigkeit von Avataren. Da wie bereits in Kapitel 2.3 ausgeführt das menschliche Denken und Handeln von der genetischen Ausstattung und durch neurobiologische Faktoren beeinflusst wird, wurden zur Steigerung des Erkenntniswerts der Studie nicht nur Verhaltensdaten erfasst, sondern auch Daten zur Gehirnaktivität der Probanden mittels fMRT gemessen. Die Analyse fokussierte dabei auf den *medialen präfrontalen Kortex*, weil bekannt ist, dass diese Hirnregion insbesondere für das *Mentalizing* bedeutsam ist – es wurde weiter oben bereits ausgeführt, dass damit die menschliche Fähigkeit bezeichnet wird, die Gedanken und Absichten anderer Menschen zu verstehen, was insbesondere in Vertrauenssituationen bedeutsam ist (Krueger et al. 2007; McCabe et al. 2001). Es wurde folgende Hypothese formuliert: Bei der Entscheidung, ob man einem anderen Akteur vertrauen soll oder nicht, aktiviert die Entscheidungsfindung den medialen präfrontalen Kortex während der Interaktion mit Menschen stärker als während der Interaktion mit Avataren. Die zugrunde liegende Argumentation kann wie folgt zusammengefasst werden (Riedl et al. 2014, S. 91):

> *„Considering that face-to-face interaction has been the major communications mode for the majority of human history, a specific trait for processing human facial information and corresponding trustworthiness prediction abilities must be part of the human genetic makeup. A number of other scholars share this view; they also regard trust and mentalizing as „evolved capacities" […] With respect to possible explanations of why such perception of human faces constitutes a basic instinct—and as a complement to the fact that trust in untrustworthy humans could have directly resulted in death—scholars have offered explanations related to the effectiveness of social functioning".*

In Bezug auf die Beantwortung der dritten Forschungsfrage wurde als Grundlage die Theorie zur Plastizität des Gehirns herangezogen. Wissenschaftliche Befunde zeigen, dass die Evolution den Menschen auch mit einem „plastischen" Gehirn ausgestattet hat. Plastizität des Gehirns bezieht sich auf seine Fähigkeit, die Struktur und Funktion zu verändern (Kolb & Whishaw 1998). Offensichtlich ist diese Eigenschaft des Gehirns vorteilhaft für das Überleben und erhöht somit die Fitness. Lernen und Gedächtnis sowie Erfahrung sind Konzepte, die untrennbar mit der Plastizität des Gehirns verbunden sind, da sie neuronale Veränderungen (z. B. Synapsenbildung) implizieren, die wiederum mit Verhaltensänderungen einhergehen (Kolb et al. 2003). Die Fähigkeit von Menschen, rasch neue Schemata im Gehirn zu entwickeln, umfasst auch Schemata, die sich auf computerbasierte Kommunikation beziehen (Kock 2004). Vor diesem Hintergrund wurde die dritte Hypothese formuliert: In einer Vertrauenssituation gibt es in Bezug auf die Lernrate der Vertrauenswürdigkeit keinen Unterschied zwischen menschlichen und Avatar-basierten Interaktionspartnern.

Um die drei Hypothesen zu testen, wurde eine Version des Vertrauensspiels (Abbildung 8) verwendet, die die Probanden im Kernspintomographen spielten. Die Probanden spielten in der Rolle des Vertrauensgebers. Die Vertrauensnehmer wurden durch Bilder von echten Menschen sowie durch Avatarbilder repräsentiert, die eine vorab programmierte Strategie spielten. In der Studie wurden jeweils vier Bilder von echten Menschen und Avataren verwendet, die jeweils einen hohen und niedrigen Grad an Vertrauenswürdigkeit aufweisen (also insgesamt 16 Bilder, der Stimulusauswahlprozess ist im Detail in Riedl et al. 2011 beschrieben).

An der fMRT-Studie nahmen 11 männliche und 8 weibliche Personen teil. Alle Teilnehmer waren gesund und hatten keine Vorgeschichte neurologischer oder psychiatrischer Erkrankungen (Selbstauskunft). Das Durchschnittsalter der Teilnehmer betrug 31,83 Jahre. Eine Teilnehmerin (weiblich) musste von weiteren Analysen ausgeschlossen werden, da sie nach der fMRT-Datenerhebung angab, dass sie Angst hatte, während sie sich im Kernspintomographen befand; sie dachte daher nicht an die experimentelle Aufgabe. Daher beträgt die finale Stichprobengröße N = 18. Alle Teilnehmer waren mit dem Internet vertraut und nutzten es seit vielen Jahren (M = 10,77, SD = 3,40). Die Studie fokussierte daher auf erfahrene Computer- und Internetnutzer.

Wie bereits angeführt spielten die Teilnehmer das Vertrauensspiel in der Rolle des Vertrauensgebers (Abbildung 8). Das Spiel ahmt eine typische Interaktion in sozialen und wirtschaftlichen Austauschbeziehungen nach, sowohl in physischen Umgebungen (Menschen-Kondition) als auch in Computer- und Online-Umgebungen (Avatar-Kondition). Den Teilnehmern wurde im Voraus mitgeteilt, dass die Vertrauensnehmer nicht auf ihre Spielstrategien reagieren würden. Es wurde jedoch betont, dass jeder Vertrauensnehmer einen spezifischen Charakter hat, der seine oder ihre Vertrauenswürdigkeit bestimmt. Die Hälfte der Vertrauensnehmer, sowohl Menschen als auch

Avatare, wurden im Voraus als vertrauenswürdig festgelegt, während die andere Hälfte als nicht vertrauenswürdig festgelegt wurde (= die vom Versuchsleiter programmierte Strategie). Vertrauenswürdige Akteure gaben in sieben von zehn Runden 30 € zurück, während nicht vertrauenswürdige Akteure nur in drei Runden 30 € zurückgaben (siehe die Auszahlungsmatrix in Abbildung 8). Abgesehen von ihrem Gesichtsausdruck hatten die Teilnehmer keine Informationen über die Vertrauensnehmer.

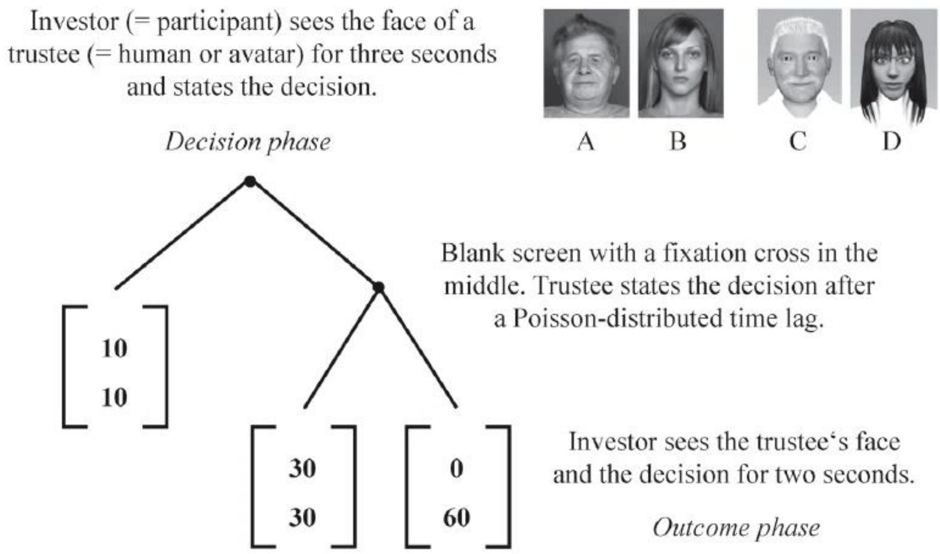


Figure 2. Structure and Payoff Matrix of the Trust Game

Notes: The upper value in the square brackets indicates the investor's payoff, and the lower value shows the trustee's payoff. Participants played against eight humans and eight avatars in the fMRI study. The human and avatar groups were matched for pre-rated trustworthiness based on facial appearance, and for behavioral trustworthiness (i.e., the number of rounds in the trust game in which money was returned). We used four trustworthy humans (see, e.g., Face A) and four untrustworthy humans (e.g., Face B), as well as four trustworthy avatars (e.g., Face C) and four untrustworthy avatars (e.g., Face D). All humans and avatars had a direct gaze, as illustrated in the four example faces. Both the human and avatar groups consisted of four males (two trustworthy and untrustworthy ones) and four females (two trustworthy and untrustworthy ones). The complete set of 16 faces used in our fMRI study is available elsewhere [101].

Abbildung 8: Auszahlungsmatrix des Vertrauensspiels und Beispielstimuli (Quelle: Riedl et al. 2014, S. 94)

In jeder Runde des Spiels wurden die Teilnehmer gefragt, ob sie ihre Geldausstattung von 10 € behalten möchten oder ob sie sie dem Vertrauensnehmer geben möchten, dessen Gesicht ihnen präsentiert wurde. Im Fall der Übergabe der 10 € an den Vertrauensnehmer wurde der Betrag um das Sechsfache multipliziert (60 €), den der Vertrauensnehmer dann entweder behalten oder aufteilen konnte (d.h. 30 € an den Teilnehmer zurückgeben). Die Teilnehmer spielten zehn Runden des Spiels mit jedem Vertrauensnehmer, wobei in jeder Runde drei Sekunden Zeit waren, um die Investitionsentscheidung zu treffen (siehe Abbildung 8, Decision Phase). Hierbei ist zu beachten, dass

Entscheidungszeiten von einer bis drei Sekunden ausreichen, um wirtschaftliche Entscheidungen in einer fMRT-Umgebung zu treffen (z. B. Winston et al. 2002). Nach einer variablen Zeit (variiert basierend auf einer Poisson-Verteilung), in der am Bildschirm mittig ein Fixationskreuz präsentiert wurde, wurde das Gesicht des Vertrauensnehmers und die Entscheidung visuell für zwei Sekunden dargestellt (siehe Abbildung 8, Outcome Phase). Vor der ersten Runde, nach der fünften Runde und nach der zehnten Runde wurden die Teilnehmer gebeten, die Vertrauenswürdigkeit des Vertrauensnehmers zu bewerten, die als Wahrscheinlichkeit operationalisiert wurde, dass der Vertrauensnehmer im Falle einer Geldübermittlung sich vertrauenswürdig verhalten würde (d.h. 30 € zurückgeben würde).

Die Verhaltensergebnisse zeigen, dass die Teilnehmer ein beträchtliches Vertrauensverhalten zeigten, unabhängig davon, ob sie mit Menschen oder Avataren spielten (durchschnittliches Vertrauen in der Menschengruppe: in 52 von 80 Spielen, SD = 11,55; durchschnittliches Vertrauen in der Avatargruppe: in 51,39 von 80 Spielen, SD = 13,57). Es gab keinen signifikanten Unterschied in der Anzahl der Entscheidungen zu vertrauen zwischen den Mensch- und Avatar-Bedingungen (t = 0,445, df = 17, p = 0,662). Die Ergebnisse der Hypothesentests stellten sich wie im Folgenden beschrieben dar.

Hypothese 1 konnte bestätigt werden. Es konnte kein signifikanter Unterschied in den Vertrauensentscheidungen der Teilnehmer beim Spielen mit vertrauenswürdigen versus nicht vertrauenswürdigen Avataren festgestellt werden (25,89 vs. 25,50, t = 0,408, df = 17, p = 0,689), jedoch wurde beim Spielen mit vertrauenswürdigen versus nicht vertrauenswürdigen Menschen ein signifikanter Unterschied festgestellt (27,61 vs. 24,39, t = 3,502, df = 17, p = 0,003).[16] Die Differenz zwischen den Differenzwerten (Menschen Δ 3.22 und Avatare Δ 0,38) ist statistisch signifikant (t = 2,536, df = 17, p = 0,021). Diese Ergebnisse deuten daher darauf hin, dass Menschen besser in der Lage sind, die Vertrauenswürdigkeit von Menschen einzuschätzen als die Vertrauenswürdigkeit von Avataren.

In der Studie lag der Fokus auf der Untersuchung der Gehirnaktivierung während der Decision Phase (siehe Abbildung 8) (Whole-Brain-Analyse). In dieser Phase mussten die Teilnehmer entscheiden, ob sie ihre anfängliche Geldausstattung von 10 € an den Vertrauensnehmer senden oder nicht. Die fMRT-Ergebnisse zeigen, dass die Aktivierung im DMPFC, rACC und VMPFC in der Menschengruppe im Vergleich zur Avatargruppe signifikant höher war (z > 3,09; Clustergröße > 25; siehe Abbildung 9). Diese drei Hirnregionen entsprechen ungefähr den Brodmann-Arealen (BA) 9, 10, 24 und 32, und sie werden als medialer frontaler Kortex anatomisch zusammengefasst, einer Hirnstruktur, die eine

[16] Die Vertrauenswürdigkeit wurde auf der Basis eines Pre-Tests bestimmt; Details dazu finden sich in Riedl et al. (2011).

bedeutende Rolle bei der neuronalen Umsetzung von Mentalisierung spielt (z. B. Amodio & Frith 2006). Dieses Ergebnis bestätigt Hypothese 2.

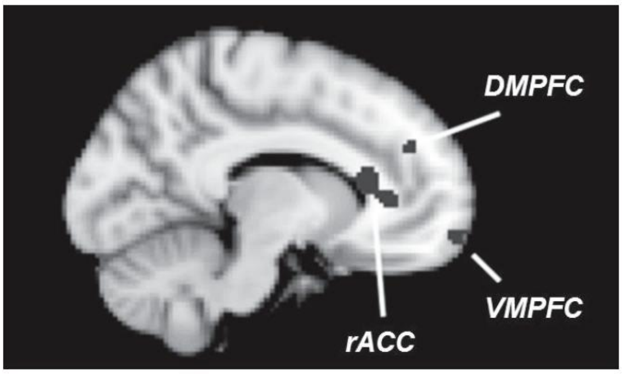


Figure 3. Brain Activation Results (Sagittal Cut)

Notes: Brain activation in the *Decision phase* (fMRI contrast: human–avatar). The reverse contrast (avatar–human) did not result in significant activation differences. The three brain areas play a significant role in the neural implementation of mentalizing. DMPFC = dorsomedial prefrontal cortex, rACC = rostral anterior cingulate cortex, VMPFC = ventromedial prefrontal cortex.

Abbildung 9: Gehirnaktivierungsunterschiede Mensch- versus Avatar-Bedingung (Quelle: Riedl et al. 2014, S. 97)

Um den Lernprozess der Teilnehmer in Bezug auf Vertrauen besser zu verstehen, insbesondere ihre Lernraten, wurde die wahrgenommene Vertrauenswürdigkeit der Vertrauensnehmer auf Rundenbasis modelliert. Die zugrunde liegende – und realitätsnahe (Rudoy & Paller 2009) – Annahme dieses Ansatzes ist, dass die wahrgenommene Vertrauenswürdigkeit eines Vertrauensnehmers auf dem Gesichtsausdruck vor Beginn der ersten Runde beruht und aufgrund seines Spielverhaltens während der folgenden Runden des Spiels angepasst wird (also ob er sich als vertrauenswürdig erweist oder nicht).

Im Regelfall wählt ein Teilnehmer die Spielstrategie, die einen höheren Erwartungswert hat. Der Erwartungswert, die anfängliche Geldausstattung von 10 € zu behalten, beträgt 10 €, weil die Wahrscheinlichkeit für diesen Gewinn per Definition 100 % beträgt (siehe Abbildung 8, linker Pfad). Der Erwartungswert, die Geldausstattung von 10 € zu investieren, entspricht dem Produkt aus der wahrgenommenen Vertrauenswürdigkeit eines Vertrauensnehmers und dem möglichen Gewinn von 30 € (siehe Abbildung 8, rechter Pfad). Während der zehn Runden der Interaktion mit einem bestimmten Vertrauensnehmer wird im Regelfall die wahrgenommene Vertrauenswürdigkeit aufgrund des Verhaltens dieses Vertrauensnehmers angepasst. In der Studie wurde dieser Anpassungsprozess mit einem Reinforcement-Learning-Modell modelliert (Rescorla & Wagner 1972). Ein solches Modell geht davon aus, dass, nachdem eine Entscheidung für eine Alternative getroffen wurde, eine erhaltene Belohnung R(t) zum Zeitpunkt t mit einem Erwartungswert EV(t) verglichen wird, wobei die Abweichung d formalisiert wird als Vorhersagefehler PE: $d(t) = R(t) - EV(t)$. PE wird somit verwendet, um EV(t) zu aktualisieren, was die Optimierung von Belohnungsvorhersagen ermöglicht. Der Einfluss eines spezifischen PE auf EV(t) in Bezug auf die nächste Vertrauensentscheidung wird durch die Lernrate bestimmt. Formal wird das Reinforcement-Learning-Modell wie folgt definiert: $EV(t) = EV(t - 1) + \alpha \cdot d(t - 1)$. Aus EV(t) ist es möglich, die subjektive Wahrscheinlichkeit für vertrauenswürdiges Verhalten zum Zeitpunkt t zu berechnen, indem EV(t) durch 30 € geteilt wird. Im verwendeten Vertrauensspiel würde ein Modell mit konstanter Lernrate (siehe Formel) annehmen, dass die wahrgenommene Vertrauenswürdigkeit gleichermaßen für vertrauenswürdiges und nicht vertrauenswürdiges Verhalten aktualisiert wird. Im Experiment wurde jeder Teilnehmer gebeten, die Vertrauenswürdigkeit jedes Vertrauensnehmers vor Beginn der ersten Runde des Vertrauensspiels zu bewerten (nur aufgrund des Gesichtsausdrucks). Diese anfängliche Bewertung haben wir als Ausgangspunkt des Lernprozesses verwendet. Der freie Parameter α wurde angepasst, indem die Summe der quadrierten Unterschiede zwischen den Modellvorhersagen und den Vertrauenswürdigkeitsbewertungen eines Teilnehmers nach der fünften und zehnten Runde minimiert wurde.

Um die dritte Hypothese zu testen, wurde analysiert, ob sich die Lernraten zwischen den Bedingungen mit Menschen und Avataren signifikant unterscheiden. Wie erwartet konnten keine signifikanten Unterschiede gefunden werden ($t = 0{,}1510$, $df = 17$, $p = 0{,}8820$). Dieses Ergebnis bestätigt Hypothese 3.

Table 2. Trustworthiness Learning Rates (Human and Avatar)

Trustworthiness learning rates	Mean (SD)
Sum of squared differences (constant learning rate, human)	0.2343 (0.1322)
Sum of squared differences (constant learning rate, avatar)	0.2773 (0.1950)
Alpha (human)	0.2139 (0.0825)
Alpha (avatar)	0.2122 (0.0916)
Correct model predictions (constant learning rate, human)	73.61% (9.40%)
Correct model predictions (constant learning rate, avatar)	71.51% (13.92%)

Tabelle 4: Lernraten der Vertrauenswürdigkeit von Menschen und Avataren (Quelle: Riedl et al. 2014, S. 99)

In einer Gesamtschau zeigen die Befunde dieser Studie, dass (1) Menschen besser in der Lage sind, die Vertrauenswürdigkeit von Menschen einzuschätzen als die von Avataren; (2) die Entscheidungsfindung darüber, ob man einem anderen Akteur vertrauen soll oder nicht, den medialen präfrontalen Kortex signifikant stärker aktiviert, wenn man mit Menschen interagiert, verglichen mit der Interaktion mit Avataren (wie beschrieben ist dieser Gehirnbereich von herausragender Bedeutung für die Vorhersage der Gedanken und Absichten anderer Individuen (Mentalisierung), was insbesondere in Vertrauenssituationen eine wichtige Fähigkeit ist); und (3) die Lernrate der Vertrauenswürdigkeit von Menschen und Avataren sich jedoch nicht signifikant voneinander unterscheidet.

3.4 Publikation 4 (Javor et al. 2016), Laborexperiment: Initiales Vertrauen von Parkinson-Patienten in Avatare: Theorie und Forschungsbefunde

In Kapitel 2.4.3 wurde bereits ausgeführt, dass Vertrauensverhalten zwar bereits mit diversen Neuropathologien wie Psychose (Lemmers-Jansen & Fett 2022) in Zusammenhang gebracht wurde, jedoch spezifisch zum Vertrauensverhalten von Parkinson-Patienten zum Publikationszeitpunkt der hier vorgestellten Studie noch keine empirischen Forschungsergebnisse vorlagen. In Anbetracht dieses Forschungsdefizits sowie der zunehmenden Bedeutung von Digitaltechnologien im Medizinbereich (z. B. bei der Diagnose und Behandlung von Krankheiten, bei der Interaktion zwischen Arzt und Patient und im Verwaltungsbereich von Krankenhäusern) wurde daher ein Verhaltensexperiment auf der Basis des Vertrauensspiels mit Parkinson-Patienten und gesunden Kontrollprobanden durchgeführt, wobei die Studienteilnehmer – wie auch schon die Probanden des Experiments in Publikation Nr. 3 – in der Rolles des Vertrauensgebers agierten und sowohl mit Menschen als auch mit Avataren spielten (man beachte, dass die Vertrauensnehmer als Bilder von Menschen und als Avatarbilder auf einem Computerbildschirm im Experiment dargeboten wurden).

In Kapitel 2.3 wurde bei den neurobiologischen Grundlagen von Vertrauen dargestellt, dass Vertrauenswahrnehmungen und -verhalten mit Aktivierung in spezifischen Hirnregionen einhergeht (Krueger & Meyer-Lindenberg 2019; Riedl & Javor 2012), insbesondere mit Aktivierung in folgenden Regionen: den Basalganglien (Belohnungs- und Lernprozesse), dem limbischen System (Risiko- und Unsicherheitswahrnehmungen) und dem Frontalkortex (Mentalizing). Interessanterweise sind es auch überwiegend diese Hirnregionen, die durch die Parkinson-Krankheit in ihrer Struktur und Funktion pathologisch verändert sind (Javor et al. 2015). Weiter wird in der Fachliteratur darüber berichtet, dass Hirnareale zur Vertrauensregulation wie medialer Frontalkortex und die Amygdala bei der Interaktion mit Avataren im Vergleich zur Interaktion mit Menschen weniger aktiviert sind (Moser et al. 2007; Riedl et al. 2014). Daraus folgt, dass Schädigungen in diesen Arealen das Vertrauensverhalten von Parkinson-Patienten gegenüber Avataren weniger beeinflussen sollte als das Vertrauensverhalten gegenüber Menschen. In Anbetracht dieses Forschungsstands wurden die folgenden zwei Hypothesen formuliert: H1: Parkinson-Patienten zeigen ein höheres initiales Vertrauen in Avatar-Gesichter als in menschliche Gesichter. H2: Parkinson-Patienten und gesunde Kontrollpersonen zeigen ein ähnliches initiales Vertrauen in Avatar-Gesichter.

Für die Studie wurden 40 Teilnehmer rekrutiert. Zwanzig Probanden waren Patienten (M 72,35 Jahre, SD 9,16, gleichgeschlechtlich verteilt), die gemäß den Kriterien der UK PDS Brain Bank mit Parkinson diagnostiziert wurden (Gibb & Lees 1988). Alle Patienten waren vor dem Experiment klinisch stabil. Die 20 gesunden Kontrollpersonen hatten ein Durchschnittsalter von 68,40 Jahren (SD 10,00, ebenfalls gleichgeschlechtlich verteilt), keine Vorgeschichte neurologischer Erkrankungen und einen neurologischen Status ohne pathologische Befunde. Es bestand kein signifikanter Unterschied zwischen den Parkinson-Patienten und den gesunden Kontrollpersonen hinsichtlich Alter, Einkommen, Bildung und Religion (alle Teilnehmer waren römisch-katholisch); zu beachten ist, dass diese Faktoren das Vertrauensverhalten beeinflussen können (Alesina & La Ferrara 2002).[17] Zudem wurden alle Patienten hinsichtlich psychiatrischer Begleiterkrankungen wie Demenz, Impulsstörungen sowie Apathie untersucht und es wurden diejenigen Personen ausgeschlossen, die pathologische Kriterien erfüllten.

Beim Empfang im Untersuchungslabor erhielten die Teilnehmer schriftliche Informationen über den Ablauf des Experiments. Danach wurde in einem persönlichen Gespräch überprüft, ob die Probanden die Anweisungen verstanden hatten, indem mehrere hypothetische Beispiele des Vertrauensspiels durchgegangen wurden. Alle Probanden beantworteten die Kontrollfragen korrekt. In der verwendeten Version des Vertrauensspiels agierten die Probanden in der Rolle des Vertrauensgebers; das Spiel wurde sitzend an einem Computer ausgeführt. Den Teilnehmern wurde gesagt, dass sie mit Menschen spielen würden, die in Form von Gesichtsfotos oder Avatar-Gesichtern präsentiert würden – in Wirklichkeit spielten die Teilnehmer jedoch mit einer vorprogrammierten Strategie, die das Verhalten eines Vertrauensnehmers simulierte. Beispielstimuli für die Mensch- und Avatargesichter und die Auszahlungsmatrix des verwendeten Vertrauensspiels sind in Abbildung 10 dargestellt.

Die Anfangsausstattung der Teilnehmer betrug 10 €, und ihnen wurde gesagt, dass sie mit echtem Geld spielen würden und dass die Vergütung für die Teilnahme am Experiment ihr Spielgewinn im Experiment sein wird. Die Teilnehmer konnten sich entscheiden, einen beliebigen Betrag an den Vertrauensnehmer zu senden (ganze Zahlen zwischen 0 € und 10 €). Die Vertrauensnehmer wurden mit 16 menschlichen Gesichtsbildern (aus bekannten Gesichtsdatenbanken; Lundqvist et al. 1998 und Oosterhof & Todorov 2008) und 16 Avatar-Gesichtern (aus Riedl et al. 2011, 2014) dargestellt. Der Auswahlprozess der menschlichen und Avatar-Stimuli stellte sicher, dass in beiden Kategorien des

[17] Die Ergebnisse zu den entsprechenden Signifikanztests finden sich in Javor et al. (2016, S. 7).

Stimulusmaterials dieselbe Vertrauenswürdigkeit und Geschlechterverteilung vorlag. Keines der menschlichen Gesichter oder Avatar-Gesichter war den Teilnehmern bekannt oder je zuvor gesehen worden. Die Teilnehmer hatten somit keine Interaktionserfahrung mit dem Stimulusmaterial. Insgesamt trafen die Teilnehmer 32 Entscheidungen im Experiment; die Reihenfolge der Gesichtspräsentation war zufällig. Jeder Teilnehmer spielte somit das Vertrauensspiel nur einmal mit demselben Vertrauensnehmer. Zudem wurde die Würfelspielaufgabe (Game-of-Dice-Task[18]) verwendet, um mögliche Verhaltensunterschiede hinsichtlich nicht-sozialen und risikoreichen Entscheidungen als Ergebniserklärung ausschließen zu können.

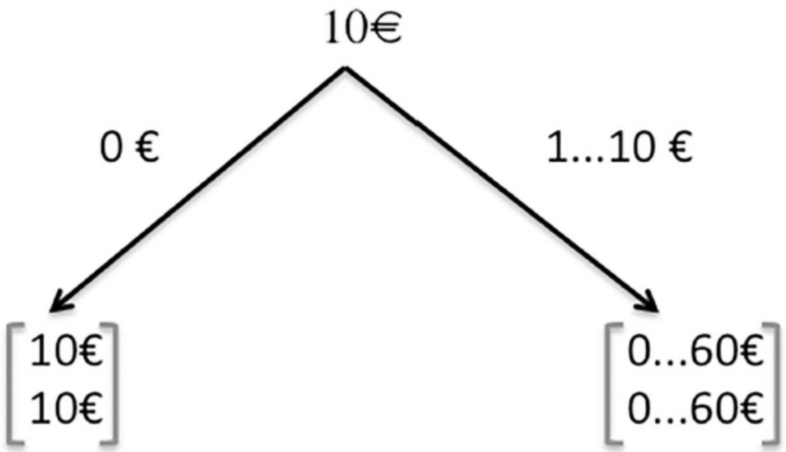


Fig 4. Illustration of our version of the trust game. Notes: The upper value in the square brackets indicates the investor's payoff (any amount between €0 and 60), the lower value the trustee's payoff depending on the investor's first move (investment of €0 or any amount between €1 and 10). When the investor is not sending any money to the trustee (€0) the payoff is €10 for both players. In any other case, the investor's payoff is dependent on the trustee's willingness to send some money back. Whereas the investor's payoff determined the overall gain of the participants in our study, the trustee's payoff was not paid out, as this role was not played by participants, but was part of the computerized experiment (see Experimental procedure).

Abbildung 10: Beispielstimuli und Auszahlungsmatrix (Quelle: Javor et al. 2016, S. 9-10)

H1 besagt, dass Parkinson-Patienten mehr initiales Vertrauen in Avatar-Gesichter zeigen als in menschliche Gesichter. Die Daten stützen diese Hypothese. Parkinson-Patienten investierten signifikant mehr in jenen Runden, die mit Avatar-Gesichtern gespielt wurden, im Vergleich zu

[18] Details dazu sind in Brand et al. (2004) dargestellt.

Runden mit menschlichen Gesichtern (M Avatar-Gesichter: 4,91 €, SD 1,55; M menschliche Gesichter: 3,43 €, SD 2,00; t = 2,62; p = 0,017). Im Gegensatz dazu unterschieden sich die durchschnittlichen Investitionsbeträge der gesunden Kontrollgruppe in Runden mit Avatar-Gesichtern im Vergleich zu menschlichen Gesichtern nicht signifikant (M Avatar: 4,94 €, SD 1,55; M menschliche Gesichter: 5,53 €, SD 1,56; t = -1,53; p = 0,141).

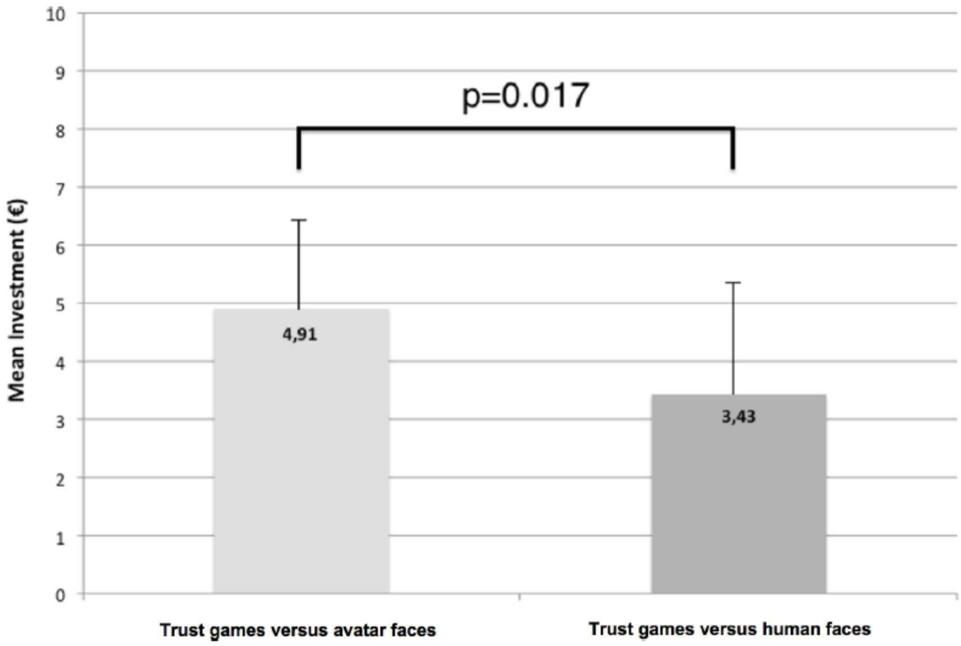


Fig 5. Results of overall mean investments in games versus avatar faces (left) and human faces (right) of the PD patients group.

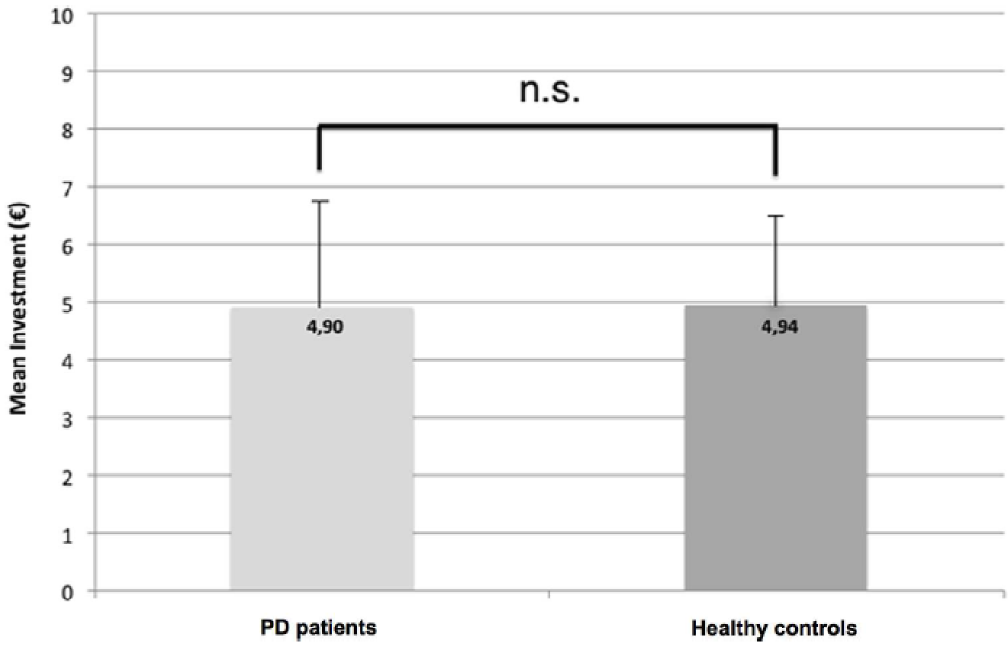


Fig 6. Results of overall mean investments in games versus avatar faces of PD patients (left) and healthy controls (right).

Abbildung 11: Grafische Ergebniszusammenfassung (Quelle: Javor et al. 2016, S. 11-12)

H2 besagt, dass Parkinson-Patienten und gesunde Kontrollpersonen ein ähnliches initiales Vertrauen in Avatar-Gesichter zeigen (und somit kein signifikanter Unterschied besteht). Die Daten stützen auch diese Hypothese. Es gab keinen signifikanten Unterschied im durchschnittlichen Investitionsbetrag zwischen der Gruppe der Parkinson-Patienten und der Kontrollgruppe in Vertrauensspielen mit Avatar-Gesichtern (M Patienten: 4,91 €, SD 1,81; M Kontrollgruppe: 4,94 €, SD 1,55; t = -0,07; p = 0,949). Parkinson-Patienten investierten hingegen im Vergleich zu gesunden Probanden signifikant weniger in den Spielen mit menschlichen Gesichtern (M Patienten: 3,43 €, SD 1,996; M Kontrollgruppe: 5,53 €, SD 1,56; t = -3,70; p = 0,001). Weiter zeigte die „Game-of-Dice-Task" einen signifikanten Unterschied im Risikoverhalten zwischen der Patientengruppe und der Kontrollgruppe, wobei eine größere Anzahl von Hochrisikoentscheidungen (aus 18 Entscheidungen) in der Patientengruppe festgestellt wurde (M Patienten: 10,2, SD 3,928; M Kontrollgruppe: 7,05, SD 3,620; t = 2,637, p = 0,012).[19] Abbildung 11 fasst die Ergebnisse zu den beiden Hypothesentests grafisch zusammen.[20]

Vertrauen ist insbesondere auch in der Arzt-Patienten-Beziehung bedeutsam, da die Befolgung von Therapiervorschlägen durch den Patienten wie die Einnahme bestimmter Medikamente durch das Vertrauen in den Arzt beeinflusst wird (z. B. Grosset et al. 2005; Thom & Campbell 1997). Es wurde gezeigt, dass Menschen mit Parkinson-Erkrankung im Vergleich zu gesunden Kontrollprobanden bei der Interaktion mit „echten" Menschen weniger vertrauen (gemessen auf der Basis des Vertrauensspiels). Dies ist eine wichtige Erklärung für den Umstand, dass Parkinson-Patienten – relativ zu anderen Patientengruppen betrachtet – eine niedrigere Rate der Befolgung von Therapievorschlägen (niedrigere Compliance) aufweisen, insbesondere betreffend die Einnahme von Medikamenten wie L-DOPA (Javor et al. 2015). Die hier vorgestellte Studie zeigt jedoch auch eine mögliche Lösung für die Problematik auf, nämlich dass es zwischen gesunden Kontrollprobanden und Parkinson-Patienten keine Unterschiede im Vertrauensverhalten gegenüber Avataren gibt. Eine mögliche Lösung des Problems der Nichtbefolgung der Einnahme von Medikamenten im Kontext von Mensch-zu-Mensch-Interaktion könnte somit sein, dass ein Avatar (über das TV-Gerät, am PC oder am Smartphone) mit einem Parkinson-Patienten kommuniziert, um die Einnahme bestimmter Medikamente zu empfehlen. Auf der Basis der hier vorgestellten Ergebnisse wäre die Vorhersage,

[19] In Javor et al. (2015) findet sich eine exakte Beschreibung der Gruppe der Parkinson-Patienten, konkret (S. 2): „All patients were mildly to moderately ill (mean UPDRS [Unified Parkinson's Disease Rating Scale] III score under treatment 19.85, SD 10.59 […] and received L-DOPA treatment (mean daily dose 337.5 mg, SD 303). Nine patients were on supplementary pramipexole medication (mean daily dose 1.85 mg, SD 1.1), four patients took ropinirole (mean daily dose 2.5 mg, SD 1.0), and two rotigotine (mean daily dose 6 mg, SD 2.83). The mean total L-DOPA equivalent daily dose (L-DOPA +dopamine agonists) was 448.782 mg (SD 371.443)." Es ist bekannt, dass die bei der Parkinson-Krankheit verabreichte Medikation zu risikoreicherem Verhalten führen kann (z. B. Imamura et al. 2008); dies ist bei der Interpretation der Befunde zu berücksichtigen.

[20] Neben der dargestellten Analyse der Daten hätte man zudem eine zweifaktorielle Varianzanalyse (ANOVA) mit einem Messwiederholungsfaktor (Mensch vs. Avatar) und einem Between-Subject-Faktor (Gesunder Mensch vs. Patient) rechnen können.

dass – im Vergleich zur Mensch-zu-Mensch-Interaktion – mehr Patienten dem Avatar-Ratschlag folgen. Zukünftige Forschung muss zeigen, ob diese Vorhersage tatsächlich bestätigt werden kann.

3.5 Publikation 5 (Riedl et al. 2017), Laborexperiment: Oxytocin, Vertrauen und Vertrauenswürdigkeit: Die moderierende Rolle von Musik

Die Rolle des neuroaktiven Hormons Oxytocin in Vertrauenssituationen wurde in Kapitel 2.3 bereits beschrieben. Insbesondere hat die Forschung auch gezeigt, dass die exogene Verabreichung von Oxytocin das Vertrauen in andere Menschen erhöhen kann (Kosfeld et al. 2005). Man könnte jedoch argumentieren, dass die externe Verabreichung von Oxytocin mit Ausnahme von Patientengruppen mit gesundheitlichen Problemen zur Linderung negativer Symptome (z. B. bei Autismus oder Angststörungen) kaum relevant ist.[21] Musik, ein allgegenwärtiger Reiz in der menschlichen Gesellschaft, kann auch zur Erhöhung von Oxytocin führen. Nilsson et al. (2009) untersuchten die Wirkung von Bettruhe mit Musik auf die Entspannung von Patienten, die gerade eine Herzoperation hinter sich hatten (Bypass-Operationen an der Herzkranzarterie und/oder Aortenklappenersatz). Insgesamt wurden 40 Patienten zufällig einer von zwei experimentellen Bedingungen zugeordnet: Musik hören während der Bettruhe (n = 20) oder nur Bettruhe (n = 20). Die Musik war „sanft, entspannend“ und wurde 30 Minuten lang mit einer Lautstärke von 50-60 dB gespielt (Nilsson, 2009, S. 2155). Die Entspannung wurde während der Bettruhe am Tag nach der Operation gemessen, indem mehrere Parameter erhoben wurden, darunter Plasma-Oxytocin und subjektive Entspannungsniveaus. Die Ergebnisse zeigen, dass in der Musikgruppe das Oxytocin im Vergleich zur Kontrollgruppe signifikant erhöht war und dass die selbstberichtete Entspannung in der Musikgruppe signifikant höher war als in der Kontrollgruppe. Diese Studie belegt somit, dass das Hören von Musik während der Bettruhe nach einer Herzoperation positive Auswirkungen auf die Entspannung haben kann, sowohl auf physiologischer (Oxytocin) als auch auf perzeptiver (selbstberichteter) Ebene. Die positive

[21] Es gibt mehrere Gründe, warum man argumentieren könnte, dass die externe Verabreichung von Oxytocin außerhalb von bestimmten Patientengruppen mit gesundheitlichen Problemen weniger relevant ist: 1. Ethik und Sicherheit: Die externe Verabreichung von Oxytocin kann ethische Bedenken hinsichtlich der Manipulation von Vertrauen und sozialen Beziehungen aufwerfen, insbesondere außerhalb klinischer Kontexte (weil es in diesen das Ziel ist, durch die Verabreichung eine negative Begleiterscheinung der Erkrankung zu mildern). Darüber hinaus ist dem Verfasser dieser Arbeit keine Studie bekannt, die die langfristigen Auswirkungen einer regelmäßigen exogenen Oxytocin-Verabreichung auf die Gesundheit und das Wohlbefinden systematisch untersucht hat. 2. Individuelle Unterschiede: Die Reaktion auf exogenes Oxytocin kann von Person zu Person variieren. Während bei einigen Menschen eine positive Wirkung auf das Vertrauen besteht, könnten bei anderen Menschen möglicherweise negative Wirkungen auf das Vertrauen beobachtbar sein. 3. Kontextabhängigkeit: Die Wirkung von Oxytocin auf Vertrauen und soziale Bindungen kann stark situationsspezifisch sein und somit vom Kontext abhängen. Was in einer bestimmten Umgebung wirksam ist, muss nicht unbedingt in anderen Kontexten ähnlich wirksam sein. Daraus folgt, dass nicht davon ausgegangen werden kann, dass die externe Verabreichung von Oxytocin universell, also kontextunabhängig, zu einer Erhöhung des Vertrauens führt. 4. Natürliche Wege zur Oxytocin-Freisetzung: Es gibt auch natürliche Wege, um die Freisetzung von Oxytocin im Körper zu fördern, wie eben z. B. das Hören von Musik. Es ist davon auszugehen, dass solche natürlichen Mechanismen im Regelfall sicherer als die externe Verabreichung von Oxytocin sind.

Wirkung von Musik auf das Wohlbefinden konnte in zahlreichen Studien auch an gesunden Probanden untermauert werden (z. B. Weinberg & Joseph 2017).

Basierend auf diesem Forschungsbefund wurde ein Laborexperiment durchgeführt. Konkret wurden die Auswirkungen von Musik auf (a) die Oxytocin-Spiegel der Vertrauensgeber (Blutprobe), (b) den Investitionsbetrag im Vertrauensspiel (Vertrauensverhalten) und (c) die Wahrnehmung der Vertrauenswürdigkeit des Vertrauensnehmers (Selbstbericht) untersucht. An der Studie nahmen 30 männliche Hochschulabsolventen und Doktoranden teil, die keine Vorgeschichte von neurologischen oder psychiatrischen Erkrankungen hatten und zum Zeitpunkt des Experiments keine Medikamente einnahmen. Es gab keinen signifikanten Altersunterschied zwischen den Gruppen (Musik: $n = 15$, $M = 27{,}2$, $SD = 5{,}3$; keine Musik: $n = 15$, $M = 26{,}1$, $SD = 6{,}2$) ($F(1, 28) = .256$, $p = .617$).

Den Teilnehmern wurde mitgeteilt, dass das Experiment ein ökonomisches Spiel umfasst und dabei Blutproben entnommen werden. Den Teilnehmern wurde auch mitgeteilt, dass (a) beide Spieler eine anfängliche Ausstattung von 10 € erhalten würden, (b) sie nur den ersten Zug des Spiels machen würden (d. h. in der Rolle des Vertrauensgebers agieren), (c) der Betrag, den sie dem anderen Spieler übermitteln, um das Sechsfache multipliziert würde, so dass der andere Spieler das Sechsfache erhalten würde, und (d) der andere Spieler dann einen beliebigen Betrag zurückzahlen könnte. Wenn ein Teilnehmer beispielsweise 10 € investierte (dies wäre die initiale Geldausstattung), würde der andere Spieler 60 € erhalten, von denen dieser andere Spieler dann einen Betrag zwischen 0 € und 70 € an den Teilnehmer zurücksenden könnte (70 € = 10 € [anfängliche Ausstattung] + 60 € [erhaltener Betrag]).

In unserer Version des Vertrauensspiels wurde der zweite Zug nicht gespielt, um Effekte wie Lernen oder Ärger auszuschließen (aufgrund als unfair wahrgenommener retournierter Geldbeträge). Die Teilnehmer trafen also nur die Investitionsentscheidung im ersten Spielschritt. Es wurde somit in der gegenständlichen Studie *initiales Vertrauen* untersucht, nicht jedoch erfahrungsbasiertes Vertrauen. Initiales Vertrauen ist das Vertrauen, das jemand in einen Fremden hat, bevor Interaktionserfahrungen mit der Person gemacht werden. Erfahrungsbasiertes Vertrauen ist hingegen Vertrauen, das durch die Interaktion mit der anderen Person entsteht. Die Vertrauensnehmer wurden auf einem Computerbildschirm in Form von (a) Gesichtsbildern (Simulation von Interaktion von Angesicht-zu-Angesicht) und (b) männlichen Spitznamen ohne weitere Informationen (Simulation von Interaktion in anonymen Internetumgebungen) präsentiert. Die Teilnehmer wurden darüber informiert, dass ihre Entschädigung für das Experiment das Ergebnis einer zufällig ausgewählten Spielrunde aus allen 40 Runden während des Experiments sein würde, tatsächlich erhielten jedoch alle Teilnehmer den gleichen Betrag (35 €). Die Gesichter für das Experiment stammten aus bekannten Datenbanken (Lundqvist et al. 1998; Oosterhof & Todorov 2008). Die Spitznamen wurden für diese Studie basierend auf vorhandenen männlichen Spitznamen aus Internet-Chatrooms erstellt. Es wurden jeweils 10 vertrauenswürdige und 10 nicht vertrauenswürdige Gesichter sowie dieselbe Anzahl von vertrauenswürdigen und nicht vertrauenswürdigen Spitznamen verwendet. Jeder Teilnehmer spielte während des Experiments 40 Runden, davon 20 mit „Gesichtern“ und 20 mit „Spitznamen“. Gesichter und Spitznamen wurden in Blöcken von jeweils 20 Runden präsentiert.

Nach der Einweisung der 30 Teilnehmer wurden sie zufällig einer der beiden Experimentalgruppen zugeteilt. Anschließend wurde ein Katheter zur Entnahme von mehreren Blutproben in den Arm eines jeden Teilnehmers eingeführt. Die erste Blutprobe (Blood Sample), BS_1, wurde 10 Minuten nach dem Einsetzen des Katheters entnommen. In der einen Gruppe lief während des Vertrauensspiels Musik, in der anderen Gruppe nicht. Die Musik, die nach der Entnahme von BS_1 begann und über Lautsprecher gespielt wurde, war Schuberts „Marche Militaire“ in D-Dur (Lautstärke 50-60 dB). Während des Experiments wurden sechs Blutproben (BS_1, BS_2, BS_3, BS_4, BS_5, BS_6) entnommen. Blutproben wurden vor der ersten Runde des Spiels (BS_1), nach jeweils 10 Runden des Spiels (BS_2, BS_3, BS_4, BS_5) und 10 Minuten nach dem Ende des Vertrauensspiels (BS_6) entnommen. Nachdem alle Runden des Vertrauensspiels gespielt waren, wurden die Teilnehmer (a) nach ihrem Alter gefragt und ob sie (b) die Musik erkannt haben, falls sie in der Musikgruppe waren, (c) ob ihnen die Musik gefallen hat und (d) ob sie das Musikstück vor dem Experiment kannten. Zusätzlich wurde (e) die Vertrauensneigung gemessen (Gefen 2000). Außerdem bewertete jeder Teilnehmer nachträglich (f) die Vertrauenswürdigkeit aller Gesichter und Spitznamen auf einer 7-Punkte-Likert-Skala. Das experimentelle Protokoll ist in Abbildung 12 dargestellt.

Blutproben wurden durch den Katheter in vorgekühlten Röhrchen gesammelt und während des gesamten Verarbeitungsvorgangs kalt gehalten. Die Analyse der Proben erfolgte am Institut für Psychologie der Universität Bonn. Alle experimentellen Sitzungen fanden zwischen 9:00 und 13:00 Uhr statt. Weitere methodische Details finden sich in Riedl et al. (2017).

Die Effekte von Musik auf das Vertrauensverhalten (Daten aus dem Vertrauensspiel), die wahrgenommene Vertrauenswürdigkeit der Vertrauensnehmer (Fragebogendaten) und Oxytocin (Hormondaten) wurden mittels dreier separater Varianzanalysen mit wiederholten Messungen (ANOVAs) und mit Musik (Musik, keine Musik) als Between-Subject-Faktor analysiert. In den jeweiligen Analysen fungierten das Vertrauensverhalten (vier Stufen: Gesichter 1-10, Gesichter 11-20, Spitznamen 1-10, Spitznamen 11-20), die wahrgenommene Vertrauenswürdigkeit (vier Stufen: Gesichter 1-10, Gesichter 11-20, Spitznamen 1-10, Spitznamen 11-20) und die Oxytocin-Werte (sechs Zeitpunkte der Messung) als Within-Subject-Faktor. Darüber hinaus wurden Gesamtscores für das Vertrauensverhalten und die Vertrauenswürdigkeit ermittelt, indem die Antworten aller 40

Spielrunden summiert wurden. Diese Werte wurden mittels ANOVAs auf signifikante Gruppenunterschiede (Musik, keine Musik) getestet.

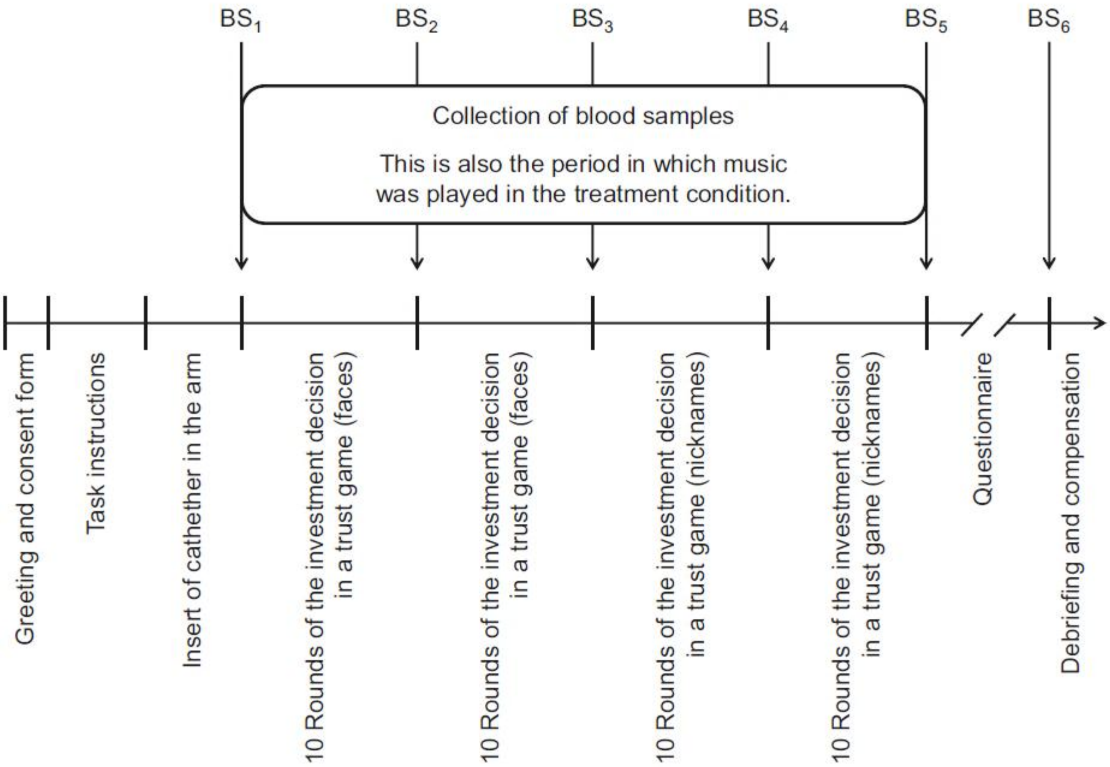


Figure 1. Experimental protocol. BS = blood sample. This protocol shows the experimental activities in sequential order (from left to right).

Abbildung 12: Experimentelles Protokoll (Quelle: Riedl et al. 2017, S. 4)

Die Ergebnisse zeigen höhere Oxytocin-Werte bei allen sechs Zeitpunkten der Messung in der Nicht-Musik-Bedingung im Vergleich zur Musik-Bedingung (siehe Abbildung 13). Die ANOVA-Ergebnisse ergaben jedoch weder signifikante Haupteffekte der Gruppe ($F(1, 28) = 1{,}93$, $p = 0{,}176$), noch der Zeit ($F(3{,}075, 86{,}090) = 0{,}692$, $p = 0{,}630$), und auch keinen signifikanten Interaktionseffekt von Zeit ×Gruppe ($F(3{,}075, 86{,}090) = 1{,}74$, $p = 0{,}164$). Die nicht signifikanten statistischen Ergebnisse könnten auf die relativ kleine Stichprobe von jeweils 15 Teilnehmern in den Gruppen zurückzuführen sein.[22]

[22] Die Effektstärke von Zeit × Gruppe kann mit einem partiellen η^2 von 5,8% angegeben werden, was auf einen merklichen Unterschied in der gemessenen Veränderung der Oxytocin-Werte zwischen den beiden experimentellen Gruppen hindeutet.

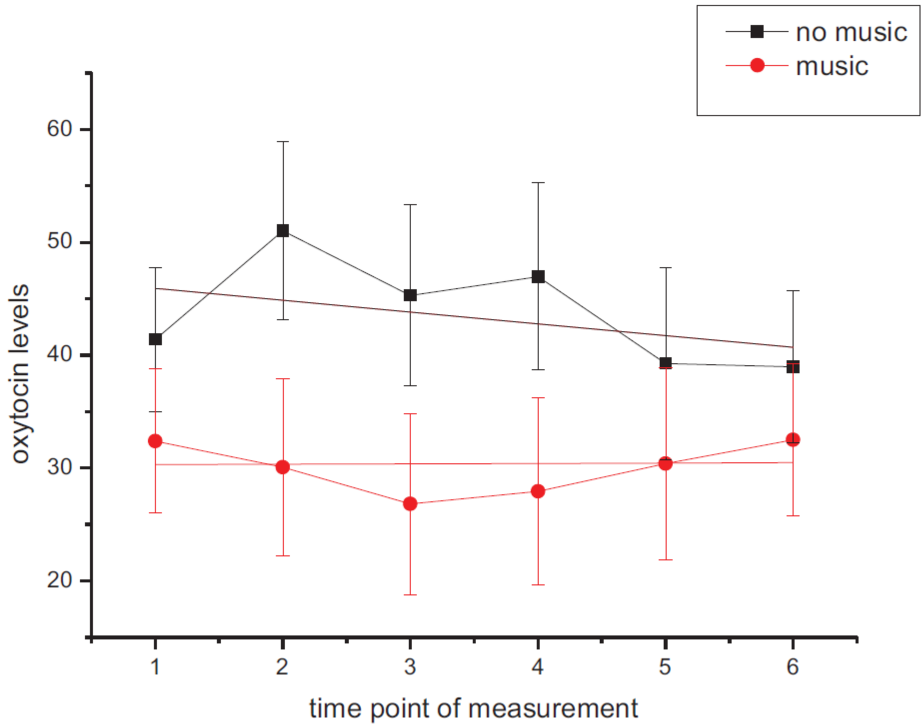


Figure 2. Oxytocin levels (pg/ml) in the treatment (music) and control (no music) groups across the six time points of measurement. Error bars indicate standard error of the means. See the online article for the color version of this figure.

Abbildung 13: Oxytocin-Ergebnisse (Quelle: Riedl et al. 2017, S. 6)

In einer zusätzlichen Analyse wurde die individuelle Veränderung des Oxytocins im Zeitablauf bestimmt, indem für jeden Teilnehmer eine lineare Anpassung basierend auf den jeweiligen sechs Oxytocin-Werten berechnet wurde. Die Regressionskoeffizienten (Beta-Gewichte [bs], Steigung der Regressionsgeraden) liefern Informationen über die Veränderungen der Oxytocin-Werte während des Experiments. Die durchschnittlichen Regressionskoeffizienten wurden zwischen den experimentellen Bedingungen verglichen (mittels ANOVA) und mit den Werten für Vertrauensverhalten und Vertrauenswürdigkeit korreliert. Basierend auf den ANOVA-Ergebnissen unterschieden sich die Steigungen in den beiden experimentellen Bedingungen nicht signifikant voneinander ($F(1, 29) = 1{,}13$, $p = 0{,}298$), obwohl eine Zunahme des Oxytocins im Zeitablauf in der Musik-Bedingung beobachtet wurde ($b = 0{,}074$), während die Oxytocin-Werte in der Nicht-Musik-Bedingung abnahmen ($b = -1{,}280$; siehe Abbildung 13). Es wurde jedoch eine signifikante Korrelationen zwischen der Steigung des Oxytocins und den Daten für Vertrauenswürdigkeit gefunden, aber nur in der Nicht-Musik-Bedingung. Konkret war die Steigung des Oxytocins signifikant korreliert mit der Vertrauenswürdigkeit von Gesichtern ($r = 0{,}550$, $p = 0{,} 023$), Spitznamen ($r = 0{,}580$, $p = 0{,}023$) und der Gesamt-Skala für Vertrauenswürdigkeit (bestehend aus Gesichtern und Spitznamen; $r = 0{,}632$, $p = 0{,}011$). Weiter gab es durchwegs positive Korrelationen zwischen der Steigung des Oxytocins und dem Vertrauensverhalten in der Nicht-Musik-Bedingung, obwohl diese Korrelationen statistisch nicht signifikant waren.

Die Ergebnisse der Studie lassen sich wie folgt zusammenfassen: Basierend auf früheren Studien, die die Auswirkungen von Oxytocin auf Vertrauen und die vertrauenssteigernden Effekte von Musik gezeigt haben, untersuchte die vorliegende explorative Studie auf der Basis des Vertrauensspiels die Auswirkungen von Musik – im Vergleich zu keiner Musik – auf endogene Oxytocin-Werte, Vertrauensverhalten und die wahrgenommene Vertrauenswürdigkeit der Vertrauensnehmer. Trotz des Fehlens eines signifikanten Effekts von Musik auf die Oxytocin-Werte, das Vertrauensverhalten und die wahrgenommene Vertrauenswürdigkeit standen individuelle Veränderungen der Oxytocin-Werte während des Vertrauensspiels in Beziehung zu den Bewertungen der wahrgenommenen Vertrauenswürdigkeit – dieser Effekt war jedoch nur in der Nicht-Musik-Bedingung vorhanden. Musik hatte somit *keinen* signifikanten Effekt auf Oxytocin, Vertrauensverhalten und die wahrgenommene Vertrauenswürdigkeit der Gegenspieler im Vertrauensspiel. Dieses Ergebnis steht im Widerspruch zu existierenden Forschungsbefunden (Anshel & Kipper 1988, Nilsson et al. 2009). Es ist jedoch unklar, ob nicht möglicherweise Unterschiede in Bezug auf die als Stimulus verwendete Musik (Anshel & Kipper 1988) sowie individuelle Unterschiede bei den rekrutierten Probanden (krank versus gesund, Nilsson et al. 2009) die Ergebnisunterschiede erklären können. Daher ist weitere Forschung wichtig.

3.6 Publikation 6 (Riedl 2022b), Review: Hängt das Vertrauen in Künstliche-Intelligenz-Systeme mit der Nutzerpersönlichkeit zusammen? Überblick über empirische Befunde und zukünftige Forschungsrichtungen

Künstliche Intelligenz (KI) bezieht sich auf Technologien, die die Ausführung von Aufgaben unterstützen, die normalerweise menschliche Intelligenz erfordern (z. B. visuelle Wahrnehmung, Spracherkennung oder Entscheidungsfindung). Beispiele für KI-Systeme sind Chatbots, Empfehlungssysteme (z. B. Amazon, Netflix), Roboter oder autonome Fahrzeuge. All diese Systeme und viele weitere sind heutzutage wichtige Phänomene in Wirtschaft und Gesellschaft. Die Entscheidung, welchem KI-System vertraut werden kann und welchem nicht, ist kritisch, da solche Systeme Aufgaben autonom ausführen und die menschliche Entscheidungsfindung beeinflussen bzw. überhaupt automatisieren. Die wachsende Bedeutung von Vertrauen in KI-Systeme geht mit einem weiteren Trend einher: dem zunehmenden Verständnis, dass die Nutzerpersönlichkeit mit dem Vertrauen zusammenhängt und somit die Akzeptanz und Nutzungsentscheidung von KI-Systemen beeinflusst.

Vor diesem Hintergrund wird im Beitrag zu Beginn ein konzeptioneller Bezugsrahmen entwickelt, der die Nutzerpersönlichkeit im Kontext von KI-Systemen beschreibt. Hierbei werden universelle Persönlichkeitsmerkmale (z. B. Big Five), spezifische Persönlichkeitsmerkmale (z. B. Vertrauensneigung), allgemeine Verhaltensneigungen (z. B. Vertrauen in ein bestimmtes KI-System) und spezifische Verhaltensweisen (z. B. Befolgung der Empfehlung eines KI-Systems in einem bestimmten Entscheidungskontext) unterschieden. Der Bezugsrahmen ist in Abbildung 14 visualisiert. Basierend auf diesem Bezugsrahmen wurde die wissenschaftliche Literatur im Rahmen einer Synopse untersucht, um folgende Forschungsfrage zu beantworten: *Hängt das Vertrauen in KI-Systeme mit der Nutzerpersönlichkeit zusammen?*

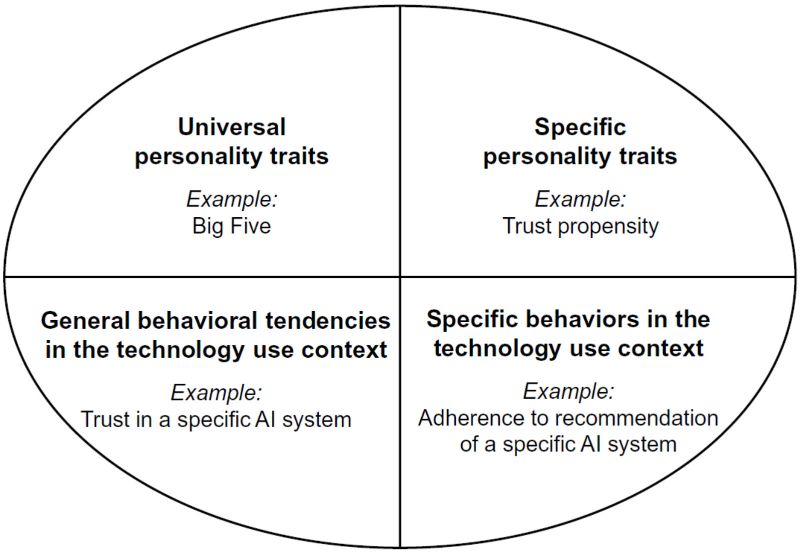


Fig. 1 Framework of user personality and trust in AI systems

Abbildung 14: Konzeptioneller Bezugsrahmen (Quelle: Riedl 2022b, S. 2025)

Auf der Basis eines umfangreichen Rechercheprozesses und rigorosen Analyseprotokolls wurden N=58 empirische Studien aus verschiedenen wissenschaftlichen Disziplinen (z. B. Informatik, Psychologie) analysiert, um den Forschungsstand zu den Zusammenhängen zwischen Persönlichkeitsmerkmalen und Vertrauen in KI-Systeme zu analysieren. Die Methodik der Literaturrecherche sowie das Analyseprotokoll sind im Detail in Riedl (2022b, S. 2025-2027) dargestellt. Im Folgenden werden die inhaltlichen Kernergebnisse der Analyse beschrieben; auf eine Darstellung deskriptiver Eigenschaften der untersuchten Artikel (z. B. verwendete Forschungsmethoden und das Land, indem die Datenerhebung stattfand) wird an dieser Stelle verzichtet (siehe dazu Riedl 2022b, S. 2027-2028).

In Bezug auf die universellen Persönlichkeitsmerkmale wurde festgestellt, dass 18 der 58 Artikel die Big Five zusammen untersucht haben. Darüber hinaus werden in 11 weiteren Artikeln einzelne Big-Five-Merkmale untersucht (Extraversion: 5, Neurotizismus: 3, Verträglichkeit: 2, Offenheit: 1, Gewissenhaftigkeit: 0). Weiter zeigen die Ergebnisse, dass andere universelle Persönlichkeitsmodelle in der Fachliteratur nur eine untergeordnete Rolle spielen. Konkret wurden in den untersuchten

Arbeiten die Myers-Briggs- und Eysenck-Modelle sowie HEXACO und NEO-PI-3 jeweils nur einmal angewendet.[23] Tabelle 5 fasst die beschriebenen Ergebnisse zusammen.

Table 2 Statistics on the personality traits

Traits	Findings	
Universal personality traits (N = 58 papers)	Big Five (together)	18
	Extraversion	5
	Neuroticism	3
	Agreeableness	2
	Openness	1
	Conscientiousness	0
	Myers-Briggs	1
	Eysenck	1
	HEXACO	1
	NEO-PI-3	1
Specific personality traits (N = 58 papers)	Trust propensity	9
	Dispositional trust in automation/machines	9
	Locus of control	7
	Attitude towards computers/robots	5
	Self-efficacy	5
	Technological innovativeness	4
	Need for cognition	3
	Perfect automation expectation	3
	Propensity to take risks	3
	Sensation seeking	3
	Need for interaction	2
	Self-esteem	2
	Other (e.g., computer anxiety, thrill seeking)	21

Tabelle 5: Anzahl der Untersuchungen verschiedener Persönlichkeitsmerkmale in der Fachliteratur zu KI-Systemen (Quelle: Riedl et al. 2022, S. 2029)

Weiter konnten 33 spezifische Persönlichkeitsmerkmale identifiziert werden. All jene Merkmale, die zumindest in zwei Artikeln untersucht wurden, sind in Tabelle 5 benannt (dies ist bei 12 Merkmalen der Fall). Zudem wurden 21 weitere spezifische Persönlichkeitsmerkmale identifiziert, die in einem Artikel untersucht wurden (z. B. Computerangst). Die bedeutsamsten spezifischen Persönlichkeitsmerkmale in der Fachliteratur zu KI-Systemen sind: Vertrauensneigung in Menschen und Vertrauensneigung in Automatisierung/Maschinen: jeweils 9 Artikel; es folgt die Kontrollüberzeugung („locus of control", definiert als Grad, in dem Menschen glauben, dass sie die Kontrolle über den Ausgang von Ereignissen haben, im Gegensatz zu äußeren Kräften, die außerhalb ihres Einflussbereichs liegen): 7 Artikel; Einstellung gegenüber Computern/Robotern und Selbstwirksamkeit: je 5 Artikel. In Tabelle 5 werden noch weitere spezifische Persönlichkeitsmerkmale benannt, die in weniger als fünf Studien untersucht wurden (von „technological innovativeness" mit vier Nennungen bis zu „self-esteem" mit zwei Nennungen).

[23] Es ist zu beachten, dass das Instrument NEO-PI-3 als eigenes Instrument klassifiziert wurde, weil es signifikant mehr Items hat (nämlich 240 Items), als die im Review identifizierten Studien, welche die Big Five mit anderen Instrumenten gemessen haben (oft ≤ 50 Items; siehe z. B. Chien et al. 2016).

In Bezug auf den *Zusammenhang zwischen den Big-Five-Merkmalen und Vertrauen in KI-Systeme* können die Kernergebnisse der Analyse wie folgt zusammengefasst werden (für Details wird auf Riedl 2022b verwiesen):

- Verträglichkeit: 9 Artikel berichten einen positiven Zusammenhang;
- Offenheit: 8 Artikel berichten einen positiven Zusammenhang;
- Extraversion: 5 Artikel berichten einen positiven Zusammenhang, 1 Artikel einen negativen Zusammenhang;
- Gewissenhaftigkeit: 3 Artikel berichten einen positiven Zusammenhang, 2 Artikel einen negativen Zusammenhang;
- Neurotizismus: 3 Artikel berichten einen negativen Zusammenhang.

In Bezug auf die spezifischen Persönlichkeitsmerkmale, die in zumindest zwei Artikeln untersucht wurden, wird zu folgenden Merkmalen ein *positiver* Zusammenhang berichtet: Vertrauensneigung in Menschen (3 Artikel), Erwartung einer Perfektion der Automatisierung (3 Artikel), Vertrauensneigung in Automation/Roboter (2 Artikel) und Selbstwirksamkeit (2 Artikel).

In Bezug auf die spezifischen Persönlichkeitsmerkmale, die in zumindest zwei Artikeln untersucht wurden, wird zu folgenden Merkmalen ein *negativer* Zusammenhang berichtet: internal „locus of control" (der Glaube, im Leben primär selbst den Ablauf der Dinge steuern zu können) (2 Artikel) und negative Einstellung gegenüber Computer/Roboter (2 Artikel).

In einer Gesamtschau der hier dargestellten Ergebnisse kann die formulierte Forschungsfrage, ob das Vertrauen in KI-Systeme mit der Nutzerpersönlichkeit zusammenhängt, mit einem „ja" beantwortet werden. Im Beitrag wird explizit auch darauf aufmerksam gemacht, dass bedeutsame Modelle zur Konzeptualisierung von Persönlichkeit bislang in der Fachliteratur zu KI-System keine Beachtung gefunden haben, obwohl ihr Erkenntniswert im gegenständlichen Kontext als hoch einzustufen ist – insofern besteht ein beträchtliches Forschungsdefizit. Insbesondere ist hier an die *Affective Neuroscience Personality Scales* (ANPS) zu denken (Reuter et al. 2017). Künftige Forschung sollte daher dieses Messinstrument verwenden, um den Einfluss der so operationalisierten Nutzerpersönlichkeit auf Vertrauen in KI-Systeme zu bestimmen. Im Gegensatz zu vielen anderen Persönlichkeitsfragebögen basiert ANPS nicht auf einem lexikalischen Ansatz; vielmehr ist dieser auf der Basis von Erkenntnissen neurowissenschaftlicher Studien entwickelt worden. Diese Studien legen nahe, dass es unterscheidbare neuronale Strukturen für die Basisemotionen SEEKING, CARE und PLAY (positive Emotionen) sowie FEAR, ANGER und SADNESS (negative Emotionen) gibt. Die Entwickler des ANPS schreiben, dass diese neuronalen Strukturen „in ähnlicher Weise nicht nur im Gehirn von Säugetieren, sondern auch tief im Gehirn von Menschen verankert sind. Unterschiede in diesen evolutionär sehr alten emotionalen Hirnarealen sind zentral, um Unterschiede in der Emotionalität und damit auch einem wesentlichen Bestandteil der Persönlichkeit zu verstehen“[24].

Im Beitrag werden weitere Möglichkeiten zur Ausrichtung der künftigen Forschung beschrieben, von denen nachfolgend noch eine weitere erläutert werden soll. Im Zuge der Erstellung der Synopse wurden auch die in der Fachliteratur verwendeten theoretischen Modelle untersucht. Der wesentliche Befund hierzu ist, dass zwei dominante theoretische Konzeptualisierungen identifiziert werden konnten: I: unabhängige Variable: Persönlichkeitsmerkmal, abhängige Variable: Vertrauen in das KI-System; II: unabhängige Variable: Persönlichkeitsmerkmal, Mediatorvariable: Vertrauen in das KI-System, abhängige Variable: Effekte wie die Absicht zur Verwendung eines KI-Systems (siehe Abbildung 15).

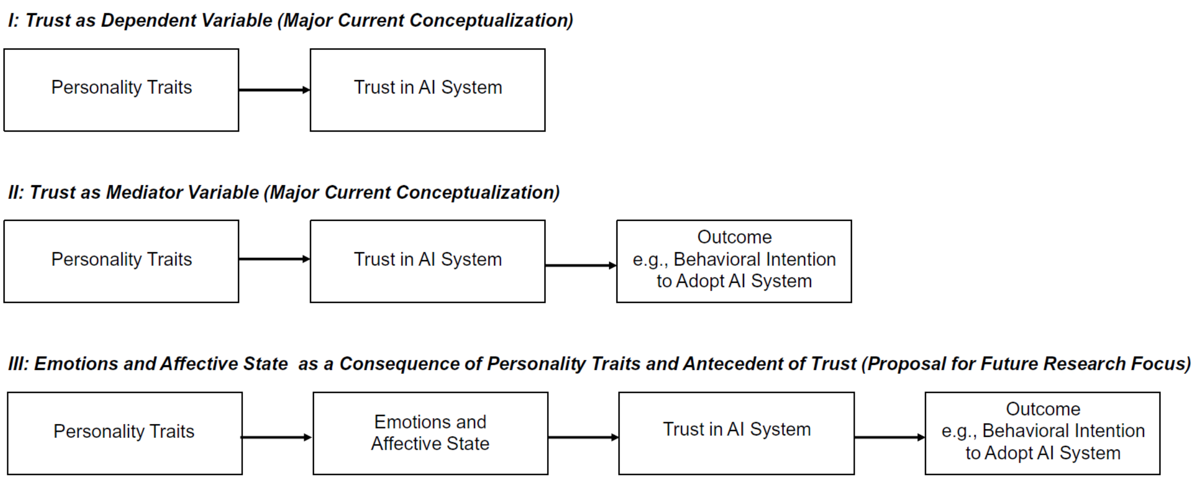


Abbildung 15: Dominierende theoretische Konzeptualisierungen und ein Vorschlag zur künftigen Ausrichtung der Forschung (Quelle: Riedl 2022b, S. 2034)

Eine mögliche Ausrichtung der zukünftigen Forschung könnte darauf fokussieren, zwischen die Variablen „Persönlichkeitsmerkmal“ und „Vertrauen in das KI-System“ eine Mediatorvariable „Emotionen und affektive Zustände“ zu integrieren, wobei Effekte wie die Absicht zur Verwendung eines KI-Systems am Ende der kausalen Kette belassen werden sollten. In Bezug auf diese mögliche Ausrichtung der künftigen Forschung wird explizit die Forderung erhoben, diese neue Mediatorvariable[25] sowohl mit Fragebögen als auch mit neurophysiologischen Verfahren zu messen. Im Beitrag steht dazu zu lesen:

[24] Quelle: https://www.hogrefe.com/at/shop/affective-neuroscience-personality-scales-anps-deutsche-version.html (Abruf am 25.03.2024).
[25] Es ist auch möglich, dass solche „Emotionen und affektiven Zustände“ situativ auftreten und somit den Charakter einer Stimmung (state) haben; dies begründet, warum sie möglicherweise auch als Moderatorvariable die Beziehung zwischen „Personality Traits“ und „Trust in AI System“ beeinflussen.

„[C]onsider measuring emotions based on neurophysiological measurement (NeuroIS). The major argument for this suggestion is that trusting beliefs in and attitudes towards AI systems are not only influenced by conscious perceptions and thoughts (which would imply the use of self-report measurement). Rather, they are also influenced by unconscious perceptions and information processing which cannot be captured based on self-reports alone […] Thus, future research should consider measuring emotions, and measurement should include both self-report and neuroscience instruments" (Riedl 2022b, S. 2035).

3.7 Publikation 7 (Riedl et al. 2024), Online-Experiment: Bevorzugen Patienten einen menschlichen Arzt, Künstliche Intelligenz oder eine Mischung und hängt diese Präferenz von der medizinischen Disziplin ab? Empirische Befunde und Implikationen für die medizinische Praxis

Es wurde eine empirische Studie im Kontext der Nutzung KI-basierter Systeme in der Medizin durchgeführt. Konkret stand hierbei die Untersuchung der Arzt-Patienten-Interaktion im Mittelpunkt. Was die Arzt-Patienten-Beziehung über fast die gesamte Zeit der Medizinhistorie geprägt hat, ist, dass die Interaktion typischerweise von Angesicht zu Angesicht stattfand und der Arzt während der Interaktion mit dem Patienten keine „intelligenten" Computersysteme verwendete. Mit dem Aufkommen von KI-Systemen wurden jedoch zwei weitere Interaktionsszenarien möglich.

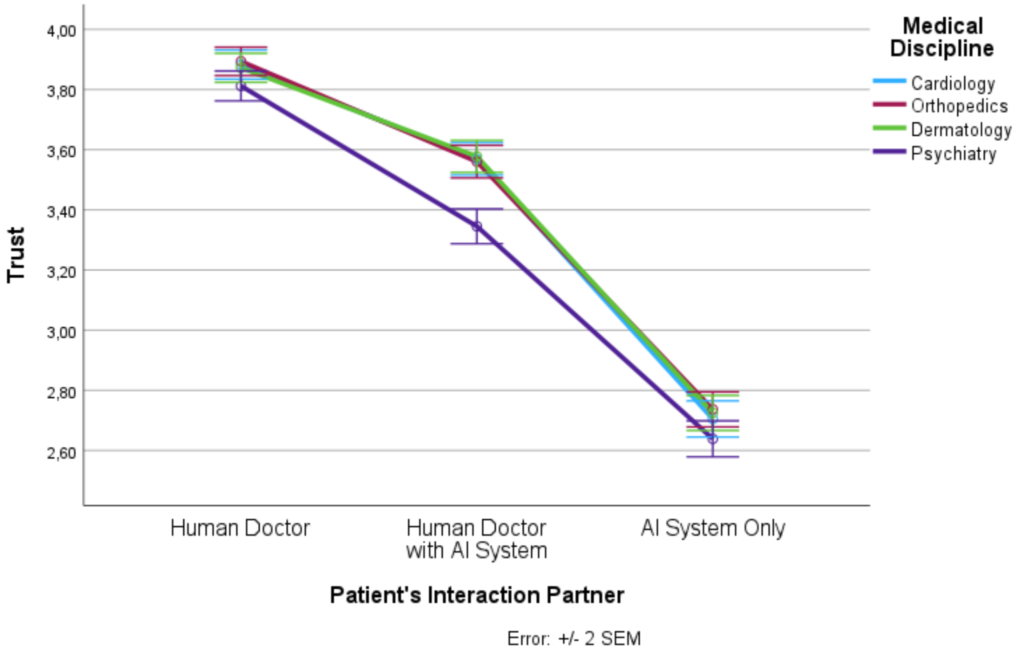
Erstens kann ein KI-System bei der Entscheidung des Arztes bezüglich Diagnose und/oder Behandlung Unterstützung leisten. Der Arzt ist Benutzer des KI-Systems, das eine Empfehlung an ihn abgibt. Der Arzt überprüft den Vorschlag, revidiert ihn bei Bedarf und trifft die endgültige Entscheidung (z. B. Jussupow et al. 2021). Somit kann ein solches KI-System als Entscheidungsunterstützungssystem betrachtet werden, das vom Arzt während seiner Interaktion mit dem Patienten verwendet wird. Zweitens ist es denkbar, dass ein KI-System den Arzt überhaupt ersetzt. Hier interagiert der Patient mit einem „intelligenten" System, aber nicht mit einem Arzt (z. B. Yokoi et al. 2021). Die Interaktion mit einem solchen „intelligenten" System manifestiert sich typischerweise als Patienten-Chatbot-Interaktion. Neben Rechenleistung und einer KI-Komponente verfügt ein solcher Chatbot über eine Benutzeroberfläche, die entweder textbasiert (der Patient tippt auf einer Tastatur und das KI-System antwortet über den Bildschirm; vergleichbar mit ChatGPT) oder sprachbasiert (gesprochene Kommunikation statt schriftlich, vergleichbar mit Sprachassistenten wie Amazons Alexa oder Apples Siri) ist (z. B. Powell 2019).

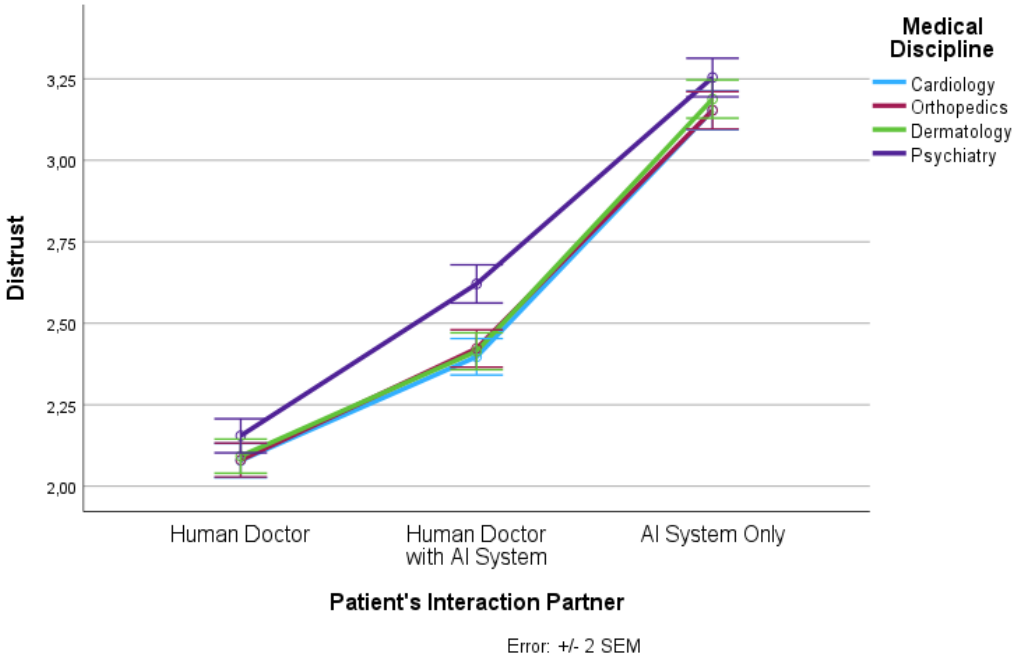
Die Studie wurde als Online-Experiment durchgeführt (N = 1.183). Die Teilnehmer wurden gebeten, sich vorzustellen, dass sie seit einiger Zeit an einer Erkrankung unbekannter Ursache leiden und daher ein neues Gesundheitszentrum aufsuchen, um sich medizinisch beraten zu lassen. Der Interaktionspartner des Patienten wurde experimentell variiert: (i) ein menschlicher Arzt, (ii) ein menschlicher Arzt mit KI-System oder (iii) nur ein KI-System (d. h. ein Chatbot). Es wurden vier medizinische Disziplinen untersucht (Kardiologie, Orthopädie, Dermatologie und Psychiatrie). Konkret wurden den Teilnehmern realitätsgetreue Beschreibungen von Patienten-Arzt-Interaktionen präsentiert und sie wurden gebeten, Fragen zu beantworten, nachdem sie mit der jeweiligen Situation konfrontiert wurden. Die Fragen bezogen sich auf sechs Variablen: Vertrauen, Misstrauen, Verletzung der Privatsphäre, Offenlegung von Informationen, Therapietreue und Zufriedenheit. Die sechs Variablen wurden ausgewählt, da sie wichtige Phänomene in der Interaktion zwischen Patient und Arzt sind; insbesondere ist es eine allgemein anerkannte Tatsache, dass die Wahrnehmung von Vertrauen und Misstrauen sowie die Wahrnehmung von Eingriffen in die Privatsphäre entscheidende Determinanten für die Offenlegung von Informationen und die Therapietreue von Patienten sind und letztendlich auch die Zufriedenheit der Patienten in ihrer Interaktion mit Ärzten beeinflussen (Dalton-Brownhaven 2020; Nagy & Sisk 2020; Pearson & Raeke 2000; Wu et al. 2022). Zudem wurden einige Variablen zum Gesundheitszustand der Befragten, zu ihrer Demographie und mehrere Variablen zu individuellen Unterschieden erhoben. In Bezug auf den Gesundheitszustand wurde anhand von „Ja“- und „Nein“-Antworten das Auftreten von Krankheiten in verschiedenen Bereichen wie Orthopädie, Dermatologie oder Innere Medizin erfasst. Darüber hinaus wurde erhoben, ob ein Teilnehmer regelmäßig Medikamente einnimmt. In Bezug auf Demographie und individuelle Unterschiede wurden Geschlecht, Alter, Bildungsstand, Persönlichkeit, Technologieeinstellung, KI-Phobie und die Bereitschaft, Menschen bzw. Maschinen zu vertrauen, erfasst.

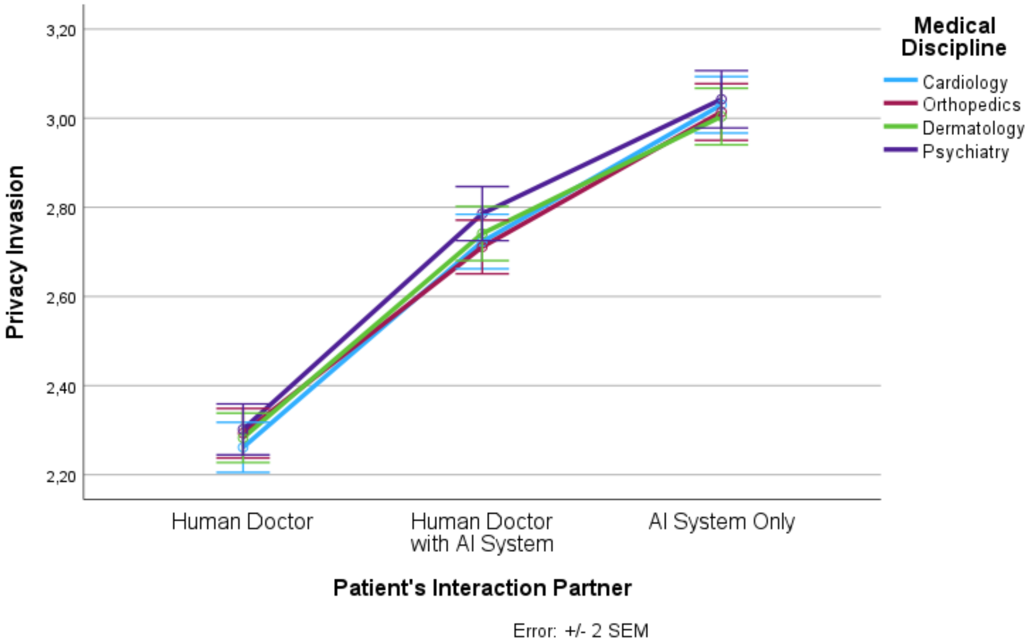
Methodisch wurde im Detail wie folgt vorgegangen: Das Online-Experiment wurde auf der Basis von fiktiven Situationsbeschreibungen („vignettes“) durchgeführt und verwendete ein Within-Subjects-Design. „Vignettes“ werden definiert als „stories about individuals and situations which make reference to important points in the study of perceptions, beliefs, and attitudes” (Hughes 1998, S. 381) und sind somit „simulations of real events” (Robert et al. 2009, S. 251). Dieser methodische Ansatz wurde zur Untersuchung medizinischer Phänomene bereits mehrfach erfolgreich eingesetzt (z. B. Gould 1996; Hughes 1998). Insgesamt wurden den Probanden 12 verschiedene Situationsbeschreibungen präsentiert: 3 Interaktionspartner des Patienten (Arzt, Arzt mit KI-System, KI-System) × 4 medizinische Disziplinen (Kardiologie, Orthopädie, Dermatologie, Psychiatrie).

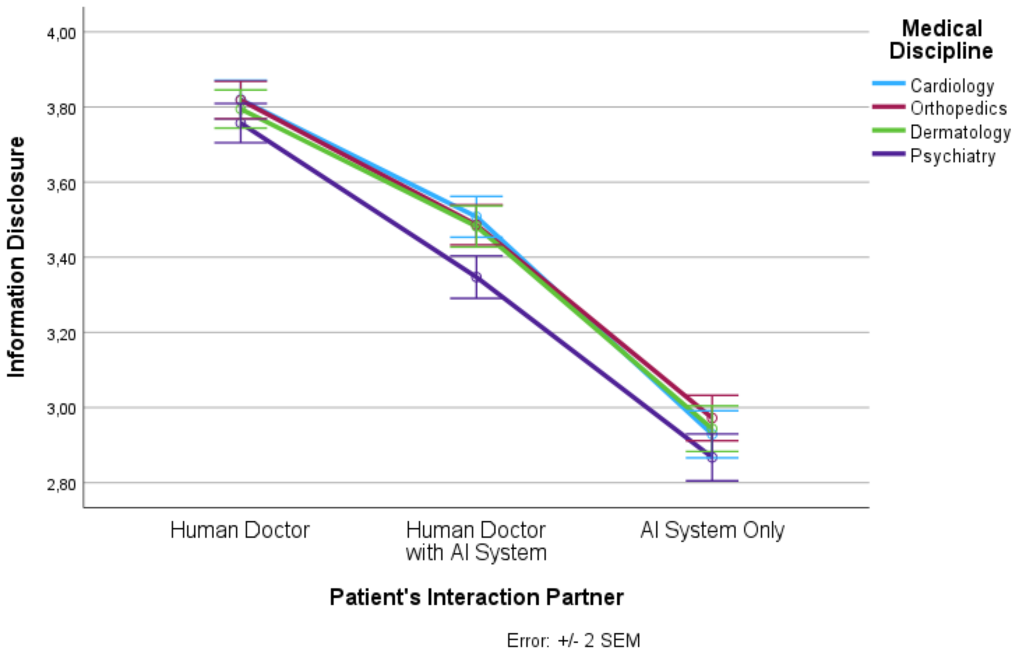
Direkt nach der Präsentation einer jeden Situation wurden die Teilnehmer gebeten, Fragen zu Vertrauen, Misstrauen, wahrgenommener Verletzung der Privatsphäre, Offenlegung von Informationen, Therapietreue und Zufriedenheit aus Patientensicht zu beantworten. Daraus folgt, dass jeder Teilnehmer die Fragen zu diesen sechs Konstrukten zwölfmal beantwortete. Schließlich wurden noch die Fragen zum Gesundheitszustand, der Demographie und zu individuellen Unterschieden gestellt. Die Daten wurden im Jahr 2022 von einem Marktforschungsunternehmen (https://www.respondi.com/) erhoben, das auf die Durchführung repräsentativer Umfragen spezialisiert ist. Die Zielpopulation der Studie waren Personen aus dem deutschsprachigen Raum (Österreich, Deutschland, Schweiz) mit einem Mindestalter von 18 Jahren. Die Teilnahmezeit betrug 29 Minuten (Median). Alle latenten Konstrukte wurden mit psychometrisch validierten Instrumenten gemessen, die auch in der vorliegenden Stichprobe eine hohe Reliabilität zeigten (Cronbach‘s Alpha > .7). Auf der Basis von Varianzanalysen (RM-ANOVA) wurde der Einfluss des Interaktionspartners des Patienten (Arzt, Arzt mit KI-System, KI-System) sowie der medizinischen Disziplin (Kardiologie, Orthopädie, Dermatologie, Psychiatrie) auf die Variablen Vertrauen, Misstrauen, Eingriff in die Privatsphäre, Offenlegung von Informationen, Therapietreue und Zufriedenheit bestimmt. Darüber hinaus wurde auf Basis von LISREL eine Strukturgleichungsmodellierung vorgenommen, um die theoretische Annahme zu überprüfen, dass Vertrauen, Misstrauen und Eingriff in die Privatsphäre einen Einfluss auf die Offenlegung von Informationen, Therapietreue und Zufriedenheit haben. Bezüglich weiterer methodischer Details wird auf Riedl et al. (2024) sowie das Supplementary Material des Beitrags verwiesen.

Die Ergebnisse in der nachfolgenden Abbildung 16 zeigen, dass Menschen einen menschlichen Arzt („Human Doctor“) bevorzugen, gefolgt von einem menschlichen Arzt mit einem KI-System („Human Doctor with AI System“) und erst dann folgt das KI-System ohne menschliche Beteiligung („AI System Only“). Konkret wurde ein signifikanter Haupteffekt des Interaktionspartners auf die Variablen Vertrauen („Trust“), Misstrauen („Distrust“), Eingriff in die Privatsphäre („Privacy Invasion“), Offenlegung von Informationen („Information Disclosure“), Therapietreue („Treatment Adherence“) und Zufriedenheit („Satisfaction“) gefunden.


Trust
4,00
3,80
3,60
3,40
3,20
3,00
2,80
2,60
Human Doctor
Human Doctor with AI System
AI System Only
Patient's Interaction Partner
Medical Discipline
Cardiology
Orthopedics
Dermatology
Psychiatry
Error: +/- 2 SEM


Distrust
3,25
3,00
2,75
2,50
2,25
2,00
Human Doctor
Human Doctor with AI System
AI System Only
Patient's Interaction Partner
Medical Discipline
Cardiology
Orthopedics
Dermatology
Psychiatry
Error: +/- 2 SEM


Medical Discipline
Cardiology
Orthopedics
Dermatology
Psychiatry
Privacy Invasion
3,20
3,00
2,80
2,60
2,40
2,20
Human Doctor
Human Doctor with AI System
AI System Only
Patient's Interaction Partner
Error: +/- 2 SEM


Medical Discipline
Cardiology
Orthopedics
Dermatology
Psychiatry
Information Disclosure
4,00
3,80
3,60
3,40
3,20
3,00
2,80
Human Doctor
Human Doctor with AI System
AI System Only
Patient's Interaction Partner
Error: +/- 2 SEM

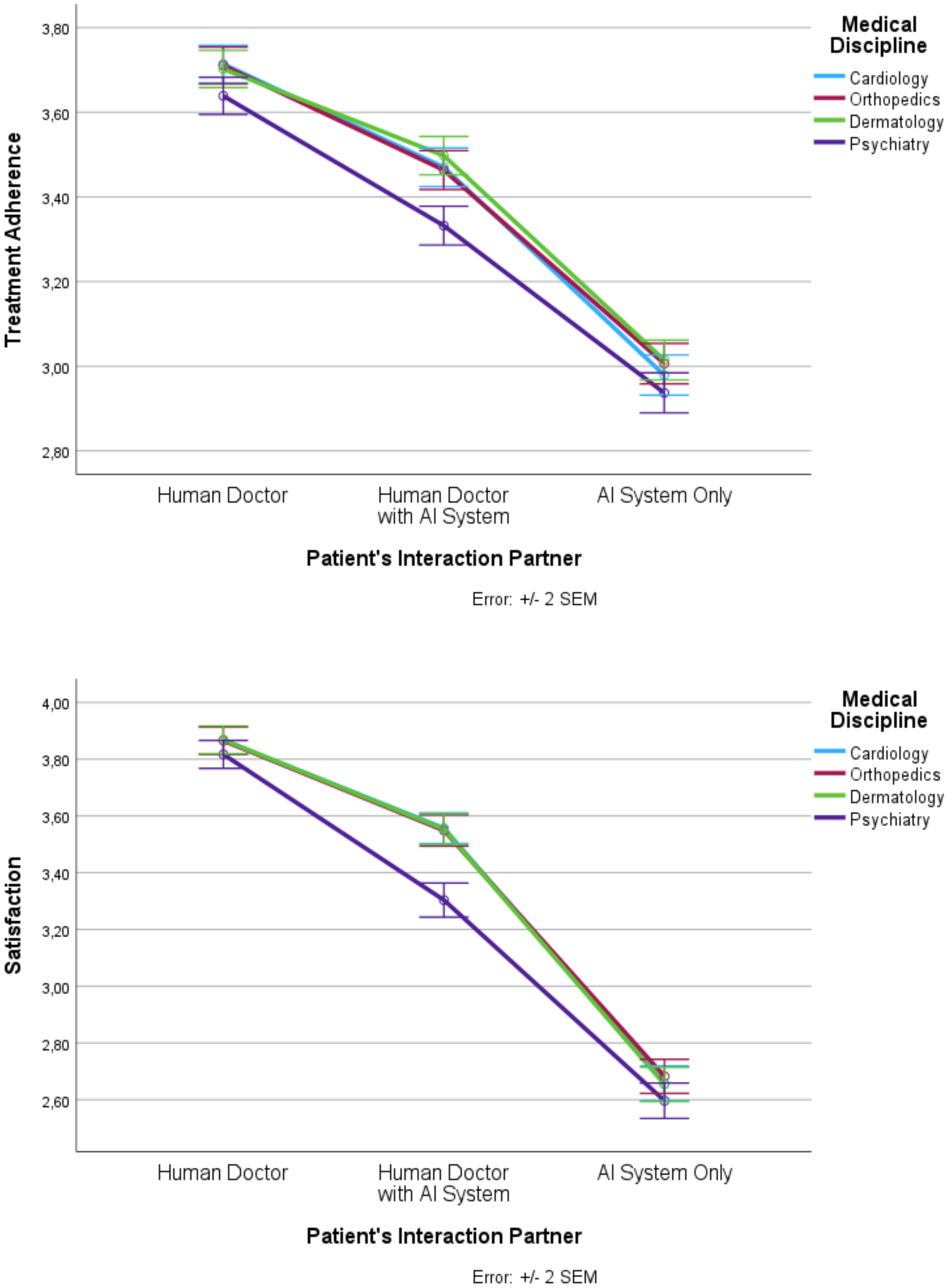


Abbildung 16: Der Einfluss des Interaktionspartners des Patienten und der medizinischen Disziplin auf sechs abhängige Variablen, die die Güte der Arzt-Patienten-Beziehung beschreiben (Quelle: Riedl et al. 2024, S. 5-8)

Bezüglich der Varianzaufklärung durch den Interaktionspartner des Patienten ergab sich das Folgende: Vertrauen (44 %), Misstrauen (40 %), Verletzung der Privatsphäre (29 %), Offenlegung von Informationen (33 %), Therapietreue (33 %) und Zufriedenheit (42 %). Zudem zeigen die Analysen, dass bei allen sechs Variablen sich die drei Manipulationsstufen („Patient's Interaction Partner,

Abbildung 16) statistisch signifikant voneinander unterscheiden (die exakten Ergebnisse der Signifikanztests finden sich in Riedl et al. 2024, S. 5-8).

In Bezug auf die Ergebnisse der Strukturgleichungsmodellierung ergaben sich die in Tabelle 6 dargestellten Ergebnisse (weitere Ergebnisse finden sich in Riedl et al. 2024, Supplementary Material). Wie bereits dargelegt wurde die theoretische Annahme zugrunde gelegt, dass Vertrauen, Misstrauen und Eingriff in die Privatsphäre als Determinanten der Offenlegung von Informationen, Therapietreue und Zufriedenheit fungieren. Die wesentliche Erkenntnis ist hierbei, dass Vertrauen einen positiven Einfluss auf die Offenlegung von Informationen, Therapietreue und Zufriedenheit ausübt; das Gesamtbild ist bei Misstrauen und Verletzung der Privatsphäre weniger klar, siehe dazu beispielsweise den positiven Zusammenhang zwischen Misstrauen und Offenlegung von Informationen, den man viel eher als negativen Zusammenhang erwarten würde. Der entscheidende Befund ist jedoch, dass Vertrauen eine zentrale Rolle einnimmt.

			Predictors		
			Trust	Distrust	Privacy invasion
Dependent variables	Information disclosure	β	1.03	0.36	−0.31
		SER	0.11	0.11	0.06
		t-value	9.23	3.36	−4.77
	Treatment adherence	β	0.45	−0.56	−0.04
		SER	0.08	0.12	0.06
		t-value	5.49	−4.72	−0.63
	Satisfaction	β	1.10	—	—
		SER	0.08	—	—
		t-value	13.96	—	—

The "—" sign in the table indicates that neither distrust nor privacy invasion has a statistically significant influence on satisfaction. Standardized regression coefficients: β. Standard errors of regression: SER.

Tabelle 6: Ergebnisse der Strukturgleichungsmodellierung (Quelle: Riedl et al. 2024, S. 9)

Weiter wurde analysiert, wie viel Varianz Vertrauen, Misstrauen und Eingriff in die Privatsphäre zusammen erklären, die Ergebnisse sind: bemerkenswerte 99 % bei der Offenlegung von Informationen und 95 % bei der Therapietreue und immer noch 24 % bei der Zufriedenheit.

Hinsichtlich der verschiedenen medizinischen Disziplinen (Kardiologie, Orthopädie, Dermatologie und Psychiatrie) wurde festgestellt, dass sich die Situation in der Psychiatrie von jener der anderen drei Disziplinen unterscheidet. Insbesondere wurden für die Psychiatrie niedrigere Werte für Vertrauen, Offenlegung von Informationen, Therapietreue und Zufriedenheit sowie höhere Werte für Misstrauen und Verletzung der Privatsphäre gefunden. Zudem war ein Interaktionseffekt zwischen Interaktionspartner des Patienten und medizinischer Disziplin zu beobachten. Die Ursache hierfür liegt darin, dass sich die sechs Variablen zwischen der Psychiatrie und den drei anderen Disziplinen in der Bedingung „Arzt mit KI-System" signifikant unterscheiden, während sie sich in den beiden anderen Bedingungen (nur Arzt, nur KI-System) nicht unterscheiden. Dieser Befund deutet darauf hin, dass Patienten die Interaktion mit einem Arzt, der von einem KI-System unterstützt wird, in der Psychiatrie mehr ablehnen als in den anderen untersuchten Disziplinen, und dieser Unterschied besteht nicht, wenn ein Patient mit einem menschlichen Arzt oder nur mit einem KI-System interagiert. Zukünftige Forschung sollte diesen interessanten Befund weiter untersuchen.

Abschließend ist der Befund bedeutsam, dass *keine* der Variablen zum Gesundheitszustand der Befragten, zu ihrer Demographie und zu individuellen Unterschieden einen maßgeblichen Einfluss auf die vorgestellten Ergebnisse hatte (die exakten Ergebnisse finden sich in Riedl et al. 2024, Supplementary Material). Dieses Ergebnis ist insbesondere in Bezug auf die Haupteffekte (siehe RM-ANOVA-Ergebnisse, Abbildung 16) erkenntnisreich, bedeutet es doch, dass die Faktoren Geschlecht, Alter, Bildungsstand, Persönlichkeit, Technologieeinstellung, KI-Phobie und die Bereitschaft, Menschen bzw. Maschinen zu vertrauen, keinen maßgeblichen Einfluss auf die Präferenzrangfolge Arzt > Arzt mit KI-System > KI-System haben. Hervorzuheben ist, dass dieses Ergebnis vor allem in Bezug auf die Persönlichkeit unerwartet ist (Oksanen et al. 2020). Persönlichkeit kann jedoch auf unterschiedliche Weise konzeptualisiert und gemessen werden. In der gegenständlichen Studie wurde Persönlichkeit auf der Grundlage der Big Five konzeptualisiert und mit dem Instrument von Gosling et al. (2003) (TIPI = Ten Item Personality Measure) gemessen. Zusätzlich zu dieser Messung universeller Persönlichkeitsmerkmale wurden vier spezifische Persönlichkeitsmerkmale erfasst (Technologieeinstellung, KI-Phobie und die Bereitschaft, Menschen bzw. Maschinen zu vertrauen). Wie in Publikation Nr. 6 dieser Dissertation dargestellt untersuchte Riedl (2022b) N = 58 empirische Artikel und identifizierte verschiedene Konzeptualisierungen universeller und spezifischer Persönlichkeitsmerkmale. Daraus folgt, dass in der zukünftigen Forschung auch andere Persönlichkeitskonzeptualisierungen und Messinstrumente verwenden werden sollten, um die hier berichteten Erkenntnisse zu replizieren, nämlich dass die Persönlichkeit die gefundenen Haupteffekte kaum beeinflusst. Dabei sollten auch im Bereich der KI-System-Forschung bislang noch gar nicht eingesetzte Instrumente verwendet werden (z. B. ANPS, Reuter et al. 2017).

In einer Gesamtschau zeigen die Ergebnisse des präsentierten Online-Experiments, dass Menschen in Arzt-Patient-Interaktionssituationen einen menschlichen Arzt bevorzugen, gefolgt von einem menschlichen Arzt mit einem KI-System, und erst dann folgt das KI-System allein. Diese Präferenz wurde anhand von sechs Faktoren untersucht, die in einer Arzt-Patienten-Beziehung eine wichtige Rolle spielen: Vertrauen, Misstrauen, Verletzung der Privatsphäre, Offenlegung von Informationen, Therapietreue und Zufriedenheit. Die wichtigste Implikation der vorgestellten Ergebnisse ist, dass der Ersatz eines menschlichen Arztes durch ein KI-System im Arzt-Patienten-Gespräch nach aktuellem Stand nicht empfohlen werden kann und daher vermieden werden sollte. Diese Schlussfolgerung steht im Einklang mit Aussagen in der Fachliteratur zum Einsatz autonomer KI-Systeme in der Medizin. Gille et al. (2020) schreiben beispielsweise dazu: „[h]owever effective, these systems offer few clues as to how they arrive at their conclusions, hence raise questions of transparency, accountability and responsibility" (S. 1). In ähnlicher Weise argumentiert Powell (2019): „[M]any medical decisions require value judgements and the doctor-patient relationship requires empathy and understanding to arrive at a shared decision, often handling large areas of uncertainty and balancing competing risks. Arguably, medicine requires wisdom more than intelligence, artificial or otherwise. Artificial intelligence therefore needs to supplement rather than replace medical professionals, and identifying the complementary positioning of artificial intelligence in medical consultation is a key challenge for the future" (S. 1). In Anbetracht der präsentierten Ergebnisse und solcher Expertenaussagen wird es spannend sein zu sehen, welche Erkenntnisse die zukünftige Forschung noch bringen wird und wie diese den Einsatz von KI-Systemen in der Arzt-Patienten-Beziehung beeinflussen werden.

4. Erkenntnisbeitrag

Dieses Kapitel gliedert sich in zwei Sub-Kapitel. Zuerst wird eine Zusammenfassung der Arbeiten und eine Darstellung der Implikationen dargeboten, danach folgen eine Darstellung der Limitationen sowie Ausblicke auf zukünftige Forschung.

4.1 Zusammenfassung, Implikationen und bisheriger wissenschaftlicher Impact

Die vorliegende Dissertation behandelt das Vertrauen in einer digitalen Welt und fokussiert dabei auf die Neurobiologie und ausgewählte individuelle Unterschiede. Insgesamt sind sieben Einzelarbeiten in die gegenständliche kumulative Dissertationsschrift eingeflossen. Wie bereits zu Beginn dieser Schrift dargelegt (siehe Abbildung 1), liegen fünf Publikationen in der Schnittmenge von *Digitale Welt*, *Vertrauen* und *Neurobiologie*, wobei zwei davon einen Fokus auf individuelle Unterschiede legen (Nr. 1, Nr. 4) und drei nicht (Nr. 2, Nr. 3, Nr. 5). Die zwei weiteren Publikationen liegen in der Schnittmenge von *Vertrauen* und *Digitale Welt*, wobei sich beide Publikationen mit individuellen Unterschieden befassen (Nr. 6, Nr. 7).

Aus dieser Einordnung folgt, dass die vorliegende Arbeit drei Themenbereiche – *Digitale Welt*, *Vertrauen* und *Neurobiologie* – integriert. Diese Themenbereiche wurden bislang kaum in der Forschung integrativ bearbeitet. Somit wird durch die vorliegende Arbeit ein wirksamer Beitrag geleistet, eine wichtige Forschungslücke zu schließen, nicht zuletzt deshalb, weil Vertrauen im Zusammenhang mit der Verwendung von Digitaltechnologien aufgrund der weltweit immer weiter ansteigenden Technologieverbreitung auch aus psychologischer Sicht zu einem äußerst bedeutsamen Thema geworden ist. Die gleichzeitige Berücksichtigung neurobiologischer Erkenntnisse sowie der spezifische Fokus auf ausgewählte individuelle Differenzen setzten weitere Akzente, die in der psychologischen Forschung seit vielen Jahrzehnten eine hohe Bedeutung haben (diese hohe Relevanz wird unter anderem auch durch die Denomination wie im Fall der Abteilung „Differentielle & Biologische Psychologie“ am Psychologieinstitut der Universität Bonn zum Ausdruck gebracht).

Die Verwendung von Digitaltechnologien kann für das Individuum, Organisationen sowie die Gesellschaft insgesamt einen Nutzen stiften, man denke hier beispielsweise an Vorteile wie direkter und oftmals kostenfreier Informationszugang, rasche(re) Kommunikationsmöglichkeiten und Produktivitätssteigerungen durch den IT-Einsatz (z. B. Heinzl et al. 2024).[26] In welchem Ausmaß und in welcher Weise diese und weitere Vorteile (z. B. Innovationspotenzial) des IT-Einsatzes jedoch wirksam werden, hängt maßgeblich vom Vertrauen in die Technologien ab. Vertrauen im technologischen Kontext besser zu verstehen ist somit wichtig und diese Dissertation hat den Anspruch, einen Erkenntnisbeitrag auf diesem Gebiet zu leisten. Die in den sieben Einzelarbeiten dargestellten Erkenntnisbeiträge befördern einerseits den wissenschaftlichen Fortschritt, sollen aber andererseits auch dazu beitragen, Vertrauensprozesse in der Praxis durch die Gestaltung von Interventionen positiv zu beeinflussen. Mit Interventionen sind hierbei in erster Linie technologische Artefakte gemeint, unter anderem eBay-Websites (Publikation 1), Avatare (Publikationen 3 und 4) und KI-Systeme (Publikationen 6 und 7).

Ein Kernthema dieser Dissertation ist die Rolle von Avataren bei der Bildung von Vertrauen. Die Ergebnisse zeigen, dass digitale Repräsentationen von Menschen das Potenzial haben, Vertrauen signifikant zu beeinflussen und soziale Interaktion zu befördern; weiter wurde festgestellt, dass es auf neurobiologischer Ebene (z. B. Aktivierung von Hirnarealen) Unterschiede in Abhängigkeit vom Interaktionspartner (Mensch, Avatar) gibt (Riedl et al. 2014). Diese Erkenntnisse sind vor allem aufgrund der zunehmenden Verwendung von Avataren im Metaverse relevant. Das Metaverse ist eine auf Virtueller Realität (VR) basierende 3D-Internetplattform für globale, digitale Gemeinschaften, in

[26] Es gibt auch Schattenseiten. Beispielsweise führen die ständige Erreichbarkeit und digitale Ablenkungen zu Überlastung und Produktivitätsrückgängen. Weiter gefährden Datenschutzbedenken und Cyberkriminalität sensible Daten und die Privatsphäre. Zudem kann exzessiver Medienkonsum zu sozialer Isolation und psychischen Problemen führen. Die ständige Verfügbarkeit von Informationen kann zudem zu Überforderung und Informationsüberflutung führen (z. B. Tarafdar et al. 2015).

denen Menschen interagieren, kommunizieren und zusammenarbeiten (Dwivedi et al. 2022). Vertrauen ist grundlegend für den Aufbau und die Erhaltung dieser im Entstehen befindlichen Gemeinschaften. Durch das Verständnis, wie Vertrauen zwischen Nutzern und ihren digitalen Repräsentationen, den Avataren, entsteht und aufrechterhalten wird, können stärkere und engagiertere Gemeinschaften im Metaverse gefördert werden.

Zudem finden im Metaverse zunehmend wirtschaftliche Aktivitäten statt, von virtuellem Handel bis hin zu Dienstleistungen (Zhang et al. 2023). Vertrauen ist eine Voraussetzung für wirtschaftliche Transaktionen, insbesondere in einem Umfeld, das durch Anonymität und Distanz gekennzeichnet ist. Die vorliegende Dissertation leistet daher einen Beitrag zu einem grundlegenden Phänomen, dem Vertrauen, das insbesondere in anonymen Umgebungen wie dem Metaverse eine entscheidende Rolle für das Funktionieren von digitalen Märkten spielt. Weiter ist zu betonen, dass im Metaverse Nutzer nicht nur mit anderen Menschen via Avatare interagieren, sondern auch mit KI-gesteuerten Avataren. Das Vertrauen in diese Technologien ist entscheidend für eine positive Nutzererfahrung (z. B. Qiu & Benbasat 2009). Durch die Erforschung der Faktoren, die das Vertrauen in KI-gesteuerte Avatare beeinflussen, können Entwickler besser gestaltete und glaubwürdige Interaktionen schaffen.

Zudem ist darauf hinzuweisen, dass die Schaffung und Nutzung von Avataren im Metaverse ethische Fragen aufwirft, insbesondere im Hinblick auf Repräsentation, Inklusivität und soziale Verantwortung. Die Forschungsarbeiten dieser Dissertation mit Bezug zu Avataren (Publikationen 3 und 4) und KI-Systemen (Publikationen 6 und 7) können dazu beitragen, ethische Richtlinien für die Gestaltung und Nutzung von Avataren zu entwickeln, die das Vertrauen der Nutzer stärken und eine positive soziale Wirkung fördern. Dabei sollten individuelle Unterschiede wie die Nutzerpersönlichkeit berücksichtigt werden.

In dieser Dissertation wurden auch medizinische Aspekte berücksichtigt (Publikationen 4 und 7). Im medizinischen Sektor könnten die Erkenntnisse dazu beitragen, digitale Repräsentationen von Ärzten zu entwickeln, die die Diagnose, Behandlung und Kommunikation mit Patienten verbessern. Dies ist vor dem Hintergrund der Forschungsergebnisse von Publikation 4 von besonderer Bedeutung für Menschen mit physischen oder psychischen Einschränkungen, die durch die neurodegenerative Parkinson-Krankheit verursacht werden. Digitaltechnologien könnten hier eine Brücke bilden, die Parkinson-Patienten hilft, effektiver mit ihrem sozialen und medizinischen Umfeld zu kommunizieren. Es gilt jedoch zu beachten, dass die Ergebnisse von Publikation 7 zeigen, dass gesunde Menschen – zumindest aktuell – eine eindeutige Präferenz für einen menschlichen Arzt in der Arzt-Patienten-Beziehung haben, wohingegen rein maschinelle KI-Systeme als Arzt sowie menschliche Ärzte, die durch KI-Systeme unterstützt werden, weniger Zustimmung finden. Da es aktuell in der medizinischen Praxis beobachtbare Tendenzen in Richtung Digitalisierung gibt, die auch KI-Systeme zur Diagnose von Krankheiten und zur Entwicklung von Therapievorschlägen miteinschließt, ist zu fordern, dass ethische und soziale Überlegungen in diesen rasanten Entwicklungsprozess einzubeziehen sind, um Systeme zu schaffen, die das menschliche Vertrauen nicht nur gewinnen (im Extremfall durch „Manipulationstechniken“ wie der anthropomorphen Gestaltung von Maschinen), sondern auch rechtfertigen (Gille et al. 2020, Powell 2019).

Zusammenfassend zeigen die durch die sieben Artikel dieser Dissertation gewonnenen Erkenntnisse, dass die Zukunft der digitalen Welt eng mit unserem Verständnis von Vertrauen, menschlichem Verhalten und seinen neurobiologischen Grundlagen sowie der Rolle von Digitaltechnologien in Gesellschaft und Wirtschaft verknüpft ist. Eine transdisziplinäre Herangehensweise (z. B. Rigolot 2020), die Benutzerzentrierung, Transparenz und ethische Überlegungen in der KI-Forschung und -Entwicklung stärker betont, ist entscheidend, um Technologien zu entwickeln, die ethisch, sozial verantwortlich und im besten Interesse der Nutzer gestaltet sind.[27]

Die hier vorgestellten Forschungsbefunde wurden bislang im wissenschaftlichen Schrifttum überwiegend so weit rezipiert, dass von einem *wahrnehmbaren wissenschaftlichen Impact* ausgegangen werden kann. Tabelle 7 fasst die dieser kumulativen Arbeit zugrunde liegenden Publikationen und ausgewählte Zitationsmetriken auf der Basis der Software „Publish or Perish 8“ (Windows GUI Edition 8.9.4538.8589, Datenbasis Google Scholar, 18.08.2024) zusammen. Die Darstellung erfolgt in chronologisch aufsteigender Reihenfolge.

Nr.	*Referenz*	*Zitationen gesamt*	*Zitationen pro Jahr Ø*
1	Riedl, R., Hubert, M., & Kenning, P. (2010). Are There Neural Gender Differences in Online Trust? An fMRI Study on the Perceived Trustworthiness of eBay Offers. *Management Information Systems Quarterly, 34*(2), 397-428.	494	35,29
2	Riedl, R., & Javor, A. (2012). The Biology of Trust: Integrating Evidence from Genetics, Endocrinology and Functional Brain Imaging. *Journal of Neuroscience, Psychology, and Economics, 5*(2), 63-91.	233	19,42
3	Riedl, R., Mohr, P., Kenning, P., Davis, F. D., & Heekeren, H. (2014). Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory. *Journal of Management Information Systems, 30*(4), 83-113.	209	20,90
4	Javor, A., Ransmayr, G., Struhal, W., & Riedl, R. (2016). Parkinson Patients’ Initial Trust in Avatars: Theory and Evidence. *PLoS ONE, 11*, 1-21.	42	5,25

Fortsetzung der Tabelle auf der nächsten Seite

[27] Der Verfasser dieser Dissertation hat gemeinsam mit Frank Krueger im Jahr 2023 die Initiative „Trandsdisciplinary Research Union for the Study of Trust“ ins Leben gerufen; weitere Informationen dazu finden sich unter: http://www.t-r-u-s-t.org/.

5	Riedl, R., Javor, A., Gefen, D., Felten, A., & Reuter, M. (2017). Oxytocin, Trust, and Trustworthiness: The Moderating Role of Music. *Journal of Neuroscience, Psychology, and Economics, 10*(1), 1-8.	16	2,29
6	Riedl, R. (2022b). Is Trust in Artificial Intelligence Systems Related to User Personality? Review of Empirical Evidence and Future Research Directions. *Electronic Markets, 32*, 2021-2051.	31	15,50
7	Riedl, R., Hogeterp, S., & Reuter, M. (2024). Do Patients Prefer a Human Doctor, Artificial Intelligence, or a Blend, and is This Preference Dependent on Medical Discipline? Empirical Evidence and Implications for Medical Practice. *Frontiers in Psychology, 15*, 1-13.	Dieser Artikel wurde am 12. August 2024 veröffentlicht. Es ist daher unzweckmäßig, Zitationsmetriken anzugeben.	

Tabelle 7: Publikationen und ausgewählte Zitationsmetriken auf der Basis der Software „Publish or Perish 8" (Windows GUI Edition 8.9.4538.8589, Datenbasis Google Scholar, 18.08.2024)

4.2 Limitationen und zukünftige Forschung

Tabelle 8 fasst die Limitationen der Publikationen dieser Dissertation auf der Basis der Angaben in den Einzelstudien zusammen (ergänzt um ein paar weitere Punkte). Im Anschluss an diese tabellarische Zusammenfassung erfolgt eine Gesamtschau zu den Limitationen über alle Veröffentlichungen hinweg, um eine Big-Picture-Perspektive zu entwickeln. Die so identifizierten Limitationen eröffnen zukünftiges Forschungspotenzial.

Nr.	*Limitationen*
1	• Stichprobengröße (N = 20, 10 Frauen und 10 Männer). • Die Ergebnisse beruhen auf einer einfachen binären Bewertung der Vertrauenswürdigkeit in einer kontrollierten Laborumgebung. • Die fMRT-Umgebung schafft eine künstliche Situation, die nicht der realen Umgebung entspricht, in der Online Shopping erfolgt. • In der Studie wurde nur ein bestimmtes Produkt (USB-Sticks) auf einem bestimmten Online-Portal (eBay) untersucht. • Die Datenanalyse umfasste keine Analysen von Hirnnetzwerken; siehe dazu Stephan et al. (2010). Zudem wurden die Effekte mit den Teilstichproben (Männer, Frauen) gerechnet, nicht aber mit dem direkten Kontrast der Geschlechter.
2	• Der dem Review zugrundliegende Recherche- und Literaturauswahlprozess ist nicht im Detail dokumentiert. Es handelt sich daher um einen *narrativen Review*, der von anderen Review-Formen, die sich durch einen höheren Grad an Systematik auszeichnen, zu unterscheiden ist; siehe dazu Paré et al. (2015).
3	• Stichprobengröße (N = 18). • Die Ergebnisse beruhen auf einer einfachen Spielaufgabe (Vertrauensspiel) in einer kontrollierten Laborumgebung. • Die fMRT-Umgebung schafft eine künstliche Situation, die nicht der realen Umgebung entspricht, in der Menschen mit Avataren interagieren. • Die Studie erfolgte auf der Basis statischer Bilder von Menschen und Avataren; insbesondere in Bezug auf Mentalizing kann dies problematisch sein, da Bewegungsabläufe (z. B. Gesten) im realen Leben als auch in der virtuellen Welt eine wichtige Informationsquelle hinsichtlich der Bestimmung der Vertrauenswürdigkeit und des Verhaltens sein können. • Avatare können verschiedene Erscheinungsformen haben. In der Studie wurde jedoch nur

	ein bestimmter Erscheinungstyp untersucht. • Die Datenanalyse umfasste keine Analysen von Hirnnetzwerken; siehe dazu Stephan et al. (2010).
4	• Das Stimulusmaterial bestand aus unbekannten menschlichen Gesichtern und unbekannten Avatar-Gesichtern. Die Ergebnisse können daher nicht auf Interaktionen mit bekannten Interaktionspartnern verallgemeinert werden. • Es wurde ausschließlich das Verhalten des Vertrauensgebers im Vertrauensspiel untersucht, nicht jedoch das Verhalten des Vertrauensnehmers. • Replikationsstudie wird aufgrund der Stichprobengröße (N = 40, 20 Parkinson-Patienten und 20 gesunde Kontrollprobanden) empfohlen. • Es sind keine direkten Rückschlüsse auf die neurologischen Ursachen des Vertrauensverhaltens der Parkinson-Patienten möglich (da nur das Vertrauensspiel zum Einsatz kam, nicht jedoch neurophysiologische Messverfahren). • Die Medikation (z. B. L-DOPA) wurde statistisch nicht kontrolliert.
5	• Stichprobengröße (N = 30, 15 Personen mit Musik, 15 Personen ohne Musik). • Es wurde nur ein bestimmtes Musikstück als Stimulus verwendet (Schuberts „Marche Militaire“ in D-Dur), obwohl bekannt ist, dass das Genre die Auswirkungen von Musik beeinflussen kann (Freeman 2000). Die Wahrnehmung von Musik ist weiter durch viele Merkmale gekennzeichnet, die in der Studie nicht systematisch variiert wurden, unter anderem Tonhöhe, Rhythmus, Klangfarbe, Tempo und Lautstärke (Levitin & Tirovolas 2009). Gomez und Danuser (2007) zeigen, dass die Wahrnehmung unterschiedlicher musikalischer Merkmale signifikante Unterschiede in physiologischen Parametern wie Atmung, Hautleitfähigkeit und Herzfrequenz hervorruft.
6	• Trotz des Umstands, dass dieser Review nach dem Kenntnisstand des Verfassers die umfangreichste Arbeit zum Zusammenhang von Nutzerpersönlichkeit und Vertrauen in KI-Systeme in der Fachliteratur ist (N = 58 identifizierte Artikel), gibt es kein formales Kriterium die Vollständigkeit in Bezug auf die berücksichtige Literatur zu belegen (die Literaturrecherche und -auswahl war jedoch umfangreich und erfolgte nach strengen Selektionskriterien). • Die Interpretation der Ergebnisse des Reviews inklusive der abgeleiteten Forschungsimplikationen ist primär subjektiv.
7	• Die Teilnehmer machten ihre Angaben nicht auf Basis einer tatsächlich stattfindenden Interaktion (mit einem Arzt, einem Arzt mit KI-System oder einem KI-System). Vielmehr beantworteten die Studienteilnehmer die Fragen zum Vertrauen usw. auf der Basis hypothetischer Situationsbeschreibungen. Robert et al. (2009) berichten jedoch über Befunde, die zeigen, dass die Ergebnisse solcher Studien durchaus hohe Aussagekraft haben können. • Die Stichprobe der vorliegenden Studie wurde im deutschsprachigen Raum (Österreich, Deutschland, Schweiz) rekrutiert. Da es Hinweise darauf gibt, dass Vertrauen mit Kultur zusammenhängt (Thanetsunthorn et al. 2019), eine Tatsache, die speziell auch für das Vertrauen in KI-Systeme gilt (Chien et al. 2016), sollten zukünftige Studien die Generalisierbarkeit der hier vorgestellten Ergebnisse prüfen, um mögliche kulturelle Unterschiede zu identifizieren. • Es wurde kein maßgeblicher Zusammenhang zwischen der Persönlichkeit und den Haupteffekten der Studie gefunden. Mit Maßgeblichkeit ist hier gemeint, dass alle Haupt- oder Interaktionseffekte der Persönlichkeitsvariablen weniger stark sind als die Effekte des Interaktionspartners des Patienten; zudem wurden ein paar wenig überraschende Effekte gefunden – z. B., dass die Neigung, Maschinen zu vertrauen, einen signifikanten Anteil der Varianz des Vertrauens eines Patienten in den Interaktionspartner erklärt. Persönlichkeit wurde auf der Grundlage der Big Five konzeptualisiert und mit einem Instrument nach Gosling et al. (2003) (TIPI = Ten Item Personality Measure) erfasst. Es kann nicht ausgeschlossen werden, dass die Verwendung eines anderen Messinstruments sowie einer anderen Konzeptualisierung von Persönlichkeit (z. B. ANPS, Reuter et al. 2017) zu anderen Ergebnissen führt.

	• In Bezug auf die Manipulation „Arzt mit KI-System“ ist anzumerken, dass die Studie ausschließlich auf die Perspektive des Patienten fokussierte. Daher war es nicht das Ziel, auch das Vertrauen des Arztes in das KI-System (zusammen mit möglichen Wahrnehmungen von Misstrauen und Eingriffen in die Privatsphäre) zu untersuchen. Es ist jedoch möglich, dass das Vertrauen des Arztes in das KI-System auch die Wahrnehmungen des Patienten und somit sein Vertrauensverhalten beeinflussen kann.

Tabelle 8: Zusammenfassung wesentlicher Limitation der Publikationen dieser Dissertation

Anmerkung: Die Nummern in der Abbildung beziehen sich auf die linke Spalte in Tabelle 1.

Tabelle 8 fasst die Limitationen der Publikationen dieser Dissertation auf der Basis der Angaben in den Einzelstudien zusammen (ergänzt um ein paar weitere Punkte). Im Anschluss an diese tabellarische Zusammenfassung erfolgt eine Gesamtschau zu den Limitationen über alle Veröffentlichungen hinweg, um eine holistische Perspektive zu entwickeln. Die so identifizierten Limitationen eröffnen zukünftiges Forschungspotenzial.

Erstens untersuchen die hier vorgestellten Arbeiten ausschließlich die Rolle des Vertrauensgebers. Zukünftige Forschung sollte daher auch die Rolle des Vertrauensnehmers fokussieren. *Zweitens* inkludiert diese Dissertation ausschließlich Laborstudien (Publikationen Nr. 1, 3, 4, 5) und eine Befragungsstudie (Publikation Nr. 7). Daraus folgt, dass zukünftige Studien auch im Feld durchgeführt werden sollten. *Drittens* sollten in Bezug auf die neurophysiologischen Messungen in zukünftigen Studien zu den hier verwendeten Methoden (2 x fMRT, 1 x Oxytocin-Messung, 1 x Vergleich von Parkinson-Patienten mit gesunden Kontrollprobanden) noch weitere Methoden eingesetzt werden – Beispiele sind EEG, fNIRS sowie EKG und Messung der elektrodermischen Aktivität; weiter sollten Methoden der Genforschung aus den Bereichen Zwillingsstudien und Molekulargenetik zum Einsatz kommen. *Viertens* sollte die zukünftige Forschung noch weitere Manifestationen der technologischen Artefakte untersuchen, da die hier verwendeten Artefakte wie eBay-Websites, Avatare, Online-Nicknames und KI-Systeme im Allgemeinen nur einen Bruchteil des möglichen Artefaktspektrums darstellen. Beispielartefakte für künftige Studien, die aktuell besonders relevant sind, wären KI-Systeme wie ChatGPT oder das Metaverse.

5. Literaturverzeichnis

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P. G., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature, 433*(7021), 68–72.

Adolphs, R., Tranel, D., & Damasio, A. (1998). The human amygdala in social judgment. *Nature, 393*, 470–474.

Alesina, A., & La Ferrara, E. (2002). Who trusts others? *Journal of Public Economics, 85*, 207–234.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268–277.

Anshel, A., & Kipper, D. A. (1988). The influence of group singing on trust and cooperation. *Journal of Music Therapy, 3*, 140–155.

APA Dictionary of Psychology. (2023). Retrieved from https://dictionary.apa.org/personality

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., ... De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*(2), 356–366.

Awad, N., & Ragowsky, A. (2008). Establishing trust in electronic commerce through online word of mouth: An examination across genders. *Journal of Management Information Systems, 24*(4), 101–121.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Baron-Cohen, S. (1999). The evolution of a theory of mind. In M. C. Corballis & S. E. G. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 261–277). Oxford University Press.

Baron-Cohen, S., Knickmeyer, R. C., & Belmonte, M. K. (2005). Sex differences in the brain: Implications for explaining autism. *Science, 310*(5749), 819–823.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58*(4), 639–650.

Bellucci, G., & Dreher, J. (2021). Trust and learning: Neurocomputational signatures of learning to trust. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 185-218). Cambridge University Press.

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems, 6*(3), 72–101.

Berg, J. E., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122–142.

Boaz, N. T., & Almquist, A. J. (2001). *Biological Anthropology: A Synthetic Approach to Human Evolution*. Prentice Hall.

Bos, P. A., Terburg, D., & Van Honk, J. (2010). Testosterone decreases trust in socially naive humans. *Proceedings of the National Academy of Sciences, 107*, 9991–9995.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8*, 539–546.

Botvinick, M. M., Nystrom, L., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature, 402*, 179–181.

Bouchard, T. J. (1994). Genes, environment, and personality. *Science, 264*(5166), 1700–1701.

Brand, M., Labudda, K., Kalbe, E., Hilker, R., Emmans, D., Fuchs, G., ... Markowitsch, H. J. (2004). Decision-making impairments in patients with Parkinson’s disease. *Behavioural Neurology, 15*(3–4), 77–85.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies, 62*(2), 161–178.

Buchan, N. R., Croson, R., & Solnick, S. J. (2008). Trust and gender: An examination of behavior and beliefs in the investment game. *Journal of Economic Behavior and Organization, 68*(3–4), 466–476.

Cacioppo, J. T., Berntson, G. G., Sheridan, J. F., & McClintock, M. K. (2000). Multilevel integrative analyses of human behavior: Social neuroscience and the complementing nature of social and biological approaches. *Psychological Bulletin, 126*(6), 829–843.

Camerer, C. F., & Weigelt, K. (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica, 56*(1), 1.

Cartwright, J. (2000). *Evolution and Human Behavior: Darwinian Perspectives on Human Nature*. MIT Press.

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., & Wallace, B. (2008). Heritability of cooperative behavior in the trust game. *Proceedings of the National Academy of Sciences of the United States of America, 105*(10), 3721–3726.

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and Big Five personality traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*, 841–845.

Coleman, J. (1990). *Foundation of Social Theory*. The Belknap Press of Harvard University Press.

Cowen, P. J., & Browning, M. (2015). What has serotonin to do with depression? *World Psychiatry, 14*(2), 158–160.

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science, 320*, 1739.

Dalton-Brown, S. (2020). The ethics of medical AI and the physician-patient relationship. *Cambridge Quarterly of Healthcare Ethics, 29*(1), 115–121.

Darwin, C. R. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray.

Database: Karolinska Directed Emotional Faces. (1998). Retrieved from http://www.emotionlab.se/resources/kdef

de Winter, R. F., Van Hemert, A. M., DeRijk, R. H., Zwinderman, K. H., Frankhuijzen-Sierevogel, A. C., Wiegant, V. M., & Goekopp, J. G. (2003). Anxious-retarded depression: Relation with plasma vasopressin and cortisol. *Neuropsychopharmacology, 28*, 140–147.

Debiec, J. (2005). Peptides of love and fear: Vasopressin and oxytocin modulate the integration of information in the amygdala. *BioEssays, 27*, 869–873.

Declerck, C. H., Boone, C., Pauwels, L., Vogt, B., & Fehr, E. (2020). A registered replication study on oxytocin and trust. *Nature Human Behaviour, 4*(6), 646–655.

De Dreu, C. K. W., Greer, L. L., Handgraaf, M. J. J., Shalvi, S., Van Kleef, G. A., Baas, M., ... Feith, S. W. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science, 328*, 1408–1411.

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological Psychiatry, 65*, 728–731.

Dinesen, P. T., Nørgaard, A. S., & Klemmensen, R. (2013). The civic personality: Personality and democratic citizenship. *Political Studies, 62*(1), 134–152.

Dirks, K. T., & De Jong, B. (2022). Trust within the workplace: A review of two waves of research and a glimpse of the third. *Annual Review of Organizational Psychology and Organizational Behavior, 9*(1), 247–276.

Dollar, D., Fisman, R., & Gatti, R. (2001). Are women really the "fairer" sex? Corruption and women in government. *Journal of Economic Behavior and Organization, 46*(4), 423–429.

Donaldson, Z. R., & Young, L. J. (2008). Oxytocin, vasopressin, and the neurogenetics of sociality. *Science, 322*(5903), 900–904.

Duke, A. A., Bègue, L., Bell, R., & Eisenlohr-Moul, T. (2013). Revisiting the serotonin-aggression relation in humans: A meta-analysis. *Psychological Bulletin, 139*(5), 1148–1172.

Dwivedi, Y. K., et al. (2022). Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 66*, 1–55.

Ebstein, R. P., Israel, S., Lerer, E., Uzefovsky, F., Shalev, I., Gritsenko, I., ... Yirmiya, N. (2009). Arginine vasopressin and oxytocin modulate human social behavior. *Annals of the New York Academy of Sciences, 1167*, 87–102.

Edelman Trust Barometer. (2023). Retrieved from https://www.edelman.com/trust/2023/trust-barometer

Lieberman, M. D., & Eisenberger, N. I. (2009). Pains and pleasures of social life. *Science, 323*, 890–891.

Eysenck, H. J. (1947). *Dimensions of personality*. In C. Kegan Paul (Ed.)

Eysenck, H. J., & Eysenck, S. B. G. (1976). Psychoticism as a dimension of personality.

Fehr, E., & Camerer, C. F. (2007). Social Neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences, 11*(10), 419–427.

Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association, 7*(2–3), 235–266.

Field, T., Grizzle, N., Scafidi, F., & Schanberg, S. (1996). Massage and relaxation therapies' effects on depressed adolescent mothers. *Adolescence, 31*, 903–911.

Fogg, B. J., & Nass, C. (1997). How users reciprocate to computers: An experiment that demonstrates behavior change. *Proceedings of the Conference on Human Factors in Computing Systems 1997*, 331–332.

Freeman, W. J. (2000). A neurobiological role of music in social bonding. In N. W. Merkur & S. Brown (Eds.), *The origins of music* (pp. 411–424). MIT Press.

Freitag, M., & Bauer, P. C. (2016). Personality traits and the propensity to trust friends and strangers. *Social Science Journal*, *53*(4), 467–476.

Friedman, B., & Millett, L. I. (1997). Reasoning about computers as moral agents: A research note. *Center for the Study of Language and Information eBooks*, 201–205.

Friedman, B., Khan, P. H., & Howe, D. C. (2000). Trust online. *Communications of the ACM, 43*(12), 34–40.

Funder, D. C. (2001). Personality. *Annual Review of Psychology, 52*(1), 197–221.

Furlan, P. M., DeMartinis, N., Schweizer, E., Rickels, K., & Lucki, I. (2001). Abnormal salivary cortisol levels in social phobic patients in response to acute psychological but not physical stress. *Biological Psychiatry, 50*, 254–259.

Garbarino, E., & Strahilevitz, M. (2003). Gender Differences in the Perveived Risk of Buying Online and the Effects of Receiving a Site Recommendation. *Journal of Business Research*, 768–775.

Gazzaniga, M. S., Ivry, R., & Mangun, G. R. (2009). Cognitive neuroscience: The biology of the mind (3rd edition). W. W. Norton.

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega: The International Journal of Management Sciences, 28*, 725–737.

Gefen, D., Benbasat, I., & Pavlou, P. (2008). A research agenda for trust in online environments. *Journal of Management Information Systems*, *24*(4), 275–286.

Gibb, W. R. G., & Lees, A. J. (1988). The relevance of the lewy body to the pathogenesis of idiopathic Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry, 51*(6), 745–752.

Gille, F., Jobin, A., & Lenca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine, 1–2*, 1–3.

Glaeser, E. L., Laibson, D., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *Quarterly Journal of Economics*, *115*(3), 811–846.

Gomez, P., & Danuser, B. (2007). Relationships between musical structure and psychophysiological measures of emotion. *Emotion, 7*, 377–387.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A Very Brief Measure of the Big Five Personality Domains. *Journal of Research in Personality, 37*, 504–528.

Gould, D. (1996). Using vignettes to collect data for nursing research studies: How valid are the findings? *Journal of Clinical Nursing, 5*(4), 207–212.

Graafland, J., & Lous, B. (2018). Income Inequality, Life Satisfaction Inequality and Trust: A Cross country panel analysis. *Journal of Happiness Studies, 20*(6), 1717–1737.

Gray, J., & McNaughton, N. (2000). The neuropsychology of anxiety. Oxford University Press.

Grosset, K. A., Bone, I., & Grosset, D. (2005). Suboptimal medication adherence in Parkinson's disease. *Movement Disorders, 20*(11), 1502–1507.

Guastella, A. J., Kenyon, A. R., Alvares, G. A., Carson, D. S., & Hickie, I. B. (2010). Intranasal arginine vasopressin enhances the encoding of happy and angry faces in humans. *Biological Psychiatry, 67*, 1220–1222.

Haas, B. W., Ishak, A., Anderson, I. W., & Filkowski, M. M. (2015). The tendency to trust is reflected in human brain structure. *NeuroImage, 107*, 175–181.

Hall, C. C., Ariss, L., & Todorov, A. (2007). The Illusion of Knowledge: When More Information Reduces Accuracy and Increases Confidence. *Organizational Behavior and Human Decision Processes, 103*(2), 277–290.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, *53*(5), 517–527.

Hanley, N. R., & Van de Kar, L. D. (2003). Serotonin and the neuroendocrine regulation of the hypothalamic-pituitary-adrenal axis in health and disease. *Vitamins and Hormones, 66*, 189–255.

Hardin, R. (2004). Trust and trustworthiness. Russell Sage Foundation.

Hauser, T. U., Eldar, E., & Dolan, R. J. (2017). Separate mesocortical and mesolimbic pathways encode effort and reward learning signals. *Proceedings of the National Academy of Sciences, 114*(35), 395–404.

Heinrichs, M., & Domes, G. (2008). Neuropeptides and social behavior. *Progress in Brain Research, 170*, 337–350.

Heinrichs, M., Baumgartner, T., Kirschbaum, C., & Ehlert, U. (2003). Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. *Biological Psychiatry, 54*, 1389–1398.

Heinrichs, M., Meinlschmidt, G., Wippich, W., Ehlert, U., & Hellhammer, D. H. (2004). Selective amnesic effects of oxytocin on human memory. *Physiology & Behavior, 83*, 31–38.

Heinzl, A., Mädche, A., & Riedl, R. (2024). Wirtschaftsinformatik – Einführung und Grundlegung (5. Auflage). Springer Gabler Verlag.

Hiraishi, K., Yamagata, S., Shikishima, C., & Ando, J. (2008). Maintenance of genetic variation in personality through control of mental mechanisms: a test of trust, extraversion, and agreeableness. *Evolution and Human Behavior, 29*(2), 79–85.

Huber, D., Pierre, V., & Ron, S. (2005). Vasopressin and oxytocin excite distinct neuronal populations in the central amygdala. *Science*, *308*, 245–248.

Hughes, R. (1998). Considering the vignette technique and its application to a study of drug injecting and HIV risk and safer behavior. *Sociology of Health and Illness, 20*(3), 381–400.

Imamura, A., Geda, Y. E., Slowinski, J., Wszolek, Z. K., Brown, L. A., & Uitti, R. J. (2008). Medications used to treat Parkinson's disease and the risk of gambling. *European Journal of Neurology, 15*(4), 350–354.

ITU. (2022). Facts and Figures – Mobile network coverage. Retrieved from https://www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-mobile-network-coverage/

Javor, A., Ransmayr, G., Struhal, W., & Riedl, R. (2016). Parkinson Patients' Initial Trust in Avatars: Theory and evidence. *PLoS ONE, 11*(11), e0165998, 1–21.

Javor, A., Riedl, R., Kirchmayr, M., Reichenberger, M., & Ransmayr, G. (2015). Trust behavior in Parkinson's Disease: Results of a Trust game experiment. *BMC Neurology, 15*, 1–7.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). The Guilford Press.

Johnson, W. (2007). Genetic and environmental influences on behavior: Capturing all the interplay. *Psychological Review, 114*(2), 423–440.

Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. *ETS Research Bulletin Series 2*, i–41.

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research, 32*(3), 713–735.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know You: Reputation and trust in a Two-Person Economic Exchange. *Science, 308*(5718), 78–83.

Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., ... & Meyer-Lindenberg, A. (2005). Oxytocin modulates neural circuitry for social cognition and fear in humans. *Journal of Neuroscience, 25*(49), 11489–11493.

Kock, N. (2004). The psychobiological model: Towards a new theory of computer-mediated communication based on Darwinian evolution. *Organization Science, 15*(3), 327–348.

Kock, N. (2005). Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward e-communication tools. *IEEE Transactions on Professional Communication*, *48*(2), 117–130.

Kock, N. (2009). Information systems theorizing based on evolutionary psychology: An interdisciplinary review and theory integration framework. *MIS Quarterly, 33*(2), 395–418.

Kock, N., & Hantula, D. A. (2005). Do we have e-collaboration genes? *International Journal of e-Collaboration, 1*(1), i–ix.

Kolb, B., Gibb, R., & Robinson, T. E. (2003). Brain plasticity and behaviour. *Current Directions in Psychological Science, 12*(1), 1–5.

Kolb, B., & Whishaw, I. Q. (1998). Brain plasticity and behaviour. *Annual Review of Psychology, 49*, 43–64.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*(7042), 673–676.

Kovács, G. L. (1986). Oxytocin and behavior. In D. Ganten & D. Pfaff (Eds.), *Neurobiology of Oxytocin* (pp. 91–128). Springer.

Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences*, *42*(2), 92–101.

Krueger, F. (Ed.). (2019). *The Neurobiology of Trust*. Cambridge University Press.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R.; Strenziok, M., Heinecke, A. & Grafman, J. (2007. Neural correlates of trust. *Proceedings of the National Academy of Sciences*, *104*(50), 20084–20089.

Lemmers-Jansen, I., & Fett, A. (2021). Trust and Psychotic Disorders: Unraveling the Dynamics of Paranoia and Disturbed Social Interaction. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 389–429). Cambridge: Cambridge University Press.

Levitin, D. J., & Tirovolas, A. K. (2009). Current advances in the cognitive neuroscience of music. *Annals of the New York Academy of Sciences, 1156*, 211–231.

Lis, S., Biermann, M., & Unoka, Z. (2021). Trust and Personality Disorders: Phenomenology, Determinants, and Therapeutical Approaches. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 430–463). Cambridge University Press.

Loup, F., Tribollet, E., Dubois-Dauphin, M., & Dreifuss, J. J. (1991). Localization of high affinity binding sites for oxytocin and vasopressin in the human brain: An autoradiographic study. *Brain Research, 555*, 220–232.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

McCabe, K. A., & Smith, V. (2000). A two person trust game played by naive and sophisticated subjects. *Proceedings of the National Academy of Sciences, 97*, 3777–3781.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835.

McCabe, K., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization, 52*(2), 267–275.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*(5), 509–516.

McCrae, R. R., & Costa, P. T. J. (1999). A Five-Factor Theory of personality. In A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). The Guilford Press.

McCrae, R. R., Costa, P. T., & Martin, T. A. (2005). The NEO–PI–3: A More Readable Revised NEO Personality Inventory. *Journal of Personality Assessment, 84*(3), 261–270.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. (2011). Trust in a specific technology. *ACM Transactions on Management Information Systems, 2*(2), 1–25.

McKnight, D. H., Choudhury, V., & Kacmar, C. J. (2002). Developing and Validating trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research, 13*(3), 334–359.

Meyer-Lindenberg, A., Domes, G., Kirsch, P., & Heinrichs, M. (2011). Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine. *Nature Reviews Neuroscience, 12*(9), 524–538.

Meyers-Levy, J. (1994). Gender Differences in Cortical Organization: Social and Biochemical Antecedents and Advertising Consequences. In E. M. Clark, T. C. Brock, & D. W. Stewart (Eds.), *Attention, Attitude, and Affect in Response to Advertising* (pp. 107–122). Lawrence Erlbaum Associates.

Meyers-Levy, J., & Maheswaran, D. (1991). Exploring Differences in Males' and Females' Processing Strategies. *Journal of Consumer Research, 18*(1), 63-70.

Mondak, J. J., & Halperin, K. (2008). A framework for the study of personality and political behaviour. *British Journal of Political Science, 38*(2), 335–362.

Moser, E., Derntl, B., Robinson, S., Fink, B., Gur, R. C., & Grammer, K. (2007). Amygdala activation at 3T in response to human and avatar facial expressions of emotions. *Journal of Neuroscience Methods, 161*(1), 126-133.

Murgatroyd, C., Wigger, A., Frank, E., Singewald, N., Bunck, M., Holsboer, F., ... Spengle, D. (2004). Impaired repression at a vasopressin promoter polymorphism underlies overexpression of vasopressin in a rat model of trait anxiety. *Journal of Neuroscience, 24*, 7762–7770.

Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *The MBTI® manual: A guide to the development and use of the Myers-Briggs type indicator*. Consulting Psychologists Press.

Nagy, M., & Sisk, B. (2020). How Will Artificial Intelligence Affect Patient-Clinician Relationships? *AMA Journal of Ethics, 22*(5), 395-400.

Nass, C., Steuer, J. S., Tauber, E., & Reeder, H. (1993). Anthropomorphism, agency, and ethopoeia: Computers as social actors. *Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems 1993*, 111-112.

Nass, C., Steuer, J., Henriksen, L., & Dryer, D. C. (1994). Machines, social attributions, and ethopoeia: performance assessments of computers subsequent to „self-" or „other-" evaluations. *International journal of human-computer studies, 40*(3), 543–559.

Nowak, M. A., & Sigmund, K. (2005). Evolution of Indirect Reciprocity. *Nature, 437*, 1291–1298.

OECD. (2017). OECD guidelines on measuring trust. Retrieved from https://www.oecd-ilibrary.org/governance/oecd-guidelines-on-measuring-trust_9789264278219-en.

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: an experimental approach to human–technology interactions online. *Frontiers in Psychology, 11*(568256), 1-13.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 11087–11092.

Oskarsson, S., Dawes, C. T., Johannesson, M., & Magnusson, P. K. E. (2012). The genetic origins of the relationship between psychological traits and social trust. *Twin Research and Human Genetics, 15*(1), 21–33.

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management, 52*, 183-199.

Patchev, V. K., & Almeida, O. F. X. (1995). Corticosteroid regulation of gene expression and binding characteristics of vasopressin receptors in the rat brain. *European Journal of Neuroscience, 7*, 1579–1583.

Pearson, S. D., & Raeke, L. H. (2000). Patients' trust in physicians: many theories, few measures, and little data. *Journal of General Internal Medicine, 15*(7), 509-513.

Petrovic, P., Kalisch, R., Singer, T., & Dolan, R. J. (2008). Oxytocin attenuates affective evaluations of conditioned faces and amygdala activity. *Journal of Neuroscience, 28*, 6607–6615.

Pfattheicher, S., & Böhm, R. (2018). Honesty-humility under threat: Self-uncertainty destroys trust among the nice guys. *Journal of Personality and Social Psychology, 114*(1), 179–194.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*(5), 729–738.

Powell, J. (2019). Trust Me, I'm a Chatbot: How Artificial Intelligence in Health Care Fails the Turing Test. *Journal of Medical Internet Research, 21*(10).

Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems, 25*(4), 145–181.

Reeves, B., & Nass, C. (1997). The media equation: how people treat computers, television, and new media like real people and places. *Choice Reviews Online, 34*(07), 34–3702.

Reimann, M., Schilke, O., & Cook, K. S. (2017). Trust is heritable, whereas distrust is not. *Proceedings of the National Academy of Sciences of the United States of America, 114*(27), 7007–7012.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). Appleton-Century-Crofts.

Reuter, M., Panksepp, J., Davis, K. L., & Montag, C. (2017). Affective neuroscience personality scales (ANPS) – Deutsche Version. Hogrefe.

Riedl, R., & Javor, A. (2012). The Biology of Trust: integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics*, *5*(2), 63–91.

Riedl, R. (2022a). Trust and Digitalization: Review of Behavioral and Neuroscience Evidence. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 54-76). Cambridge University Press.

Riedl, R. (2022b). Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electronic Markets, 32*(4), 2021–2051.

Riedl, R., Hubert, M., & Kenning, P. (2010). Are there neural gender differences in online trust? an fMRI study on the perceived trustworthiness of eBay offers. *Management Information Systems Quarterly, 34*(2), 397.

Riedl, R., Javor, A., Gefen, D., Felten, A., & Reuter, M. (2017). Oxytocin, trust, and trustworthiness: The moderating role of music. *Journal of Neuroscience, Psychology, and Economics, 10*(1), 1–8.

Riedl, R., Mohr, P. N. C., Kenning, P., Davis, F. D., & Heekeren, H. R. (2014). Trusting Humans and Avatars: A brain imaging study based on Evolution Theory. *Journal of Management Information Systems, 30*(4), 83–114.

Riedl, R., Mohr, P. N. C., Kenning, P. H., Davis, F. D., & Heekeren, H. R. (2011). Trusting humans and avatars: Behavioral and neural evidence. *Proceedings of the 32nd International Conference on Information Systems*. Association for Information Systems, 1–23.

Riedl, R., Hogeterp, S., & Reuter, M. (2024). Do Patients Prefer a Human Doctor, Artificial Intelligence, or a Blend, and is This Preference Dependent on Medical Discipline? Empirical Evidence and Implications for Medical Practice. *Frontiers in Psychology*, *15*, 1-13.

Righetti, F., & Finkenauer, C. (2011). If you are able to control yourself, I will trust you: The role of perceived self-control in interpersonal trust. *Journal of Personality and Social Psychology, 100*(5), 874–886.

Rigolot, C. (2020). Transdisciplinarity as a discipline and a way of being: complementarities and creative tensions. *Humanities and Social Sciences Communications, 7*, 1–5.

Robert, L. P., Dennis, A. R., & Hung, Y. C. (2009). Individual swift trust and knowledge trust in face-to-face and virtual team members. *Journal of Management Information Systems*, *26*(2), 241–279.

Rodgers, S., & Harris, M. A. (2003). Gender and e-Commerce: An Exploratory Study. *Journal of Advertising Research, 43*(3), 322–329.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. F. (1998). Not so different after all: a Cross-Discipline view of trust. *Academy of Management Review, 23*(3), 393–404.

Rudoy, J. D., & Paller, K. A. (2009). Who can you trust? Behavioral and neural differences between perceptual and memory-based influences. *Frontiers in Human Neuroscience, 3*, 1–6.

Saef, R., Porter, C. M., Woo, S. E., & Wiese, C. W. (2019). Getting off on the right foot: The role of openness to experience in fostering initial trust between culturally dissimilar partners. *Journal of Research in Personality, 79*, 176–187.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755–1758.

Schilke, O., Reimann, M., & Cook, K. S. (2021). Trust in social relations. *Annual Review of Sociology, 47*(1), 239–259.

Seppänen, R., Blomqvist, K., & Sundqvist, S. (2007). Measuring inter-organizational trust—A critical review of the empirical research in 1990–2003. *Industrial Marketing Management, 36*(2), 249–265.

Sheehan, K. B. (1999). An investigation of gender differences in on-line privacy concerns and resultant behaviors. *Journal of Interactive Marketing*, *13*(4), 24–38.

Shou, Q., Nishina, K., & Takagishi, H. (2021). Trust and Genetics: Genetic Basis of Trust Behavior and Trust Attitude. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 369–386). Cambridge University Press.

Sijtsma, H., & Krabbendam, L. (2021). Trust and Demographics: Age and Gender Differences in Trust and Reciprocity Behavior. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 269–292). Cambridge University Press.

Simpson, J. A., & Vieth, G. (2022). Trust and psychology: Psychological theories and principles of interpersonal trust. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 15–35). Cambridge University Press.

Simpson, T. W. (2012). What is trust? *Pacific Philosophical Quarterly, 93*, 550–569.

Skuse, D. H., & Gallagher, L. (2011). Genetic influences on social cognition. *Pediatric Research, 69*, 85R–91R.

Snijders, C., & Keren, G. (1999). Determinants of Trust. In D. Budescu, I. Erev, & R. Zwick (Eds.), *Games and Human Behavior* (pp. 355–386). Lawrence Erlbaum Associates.

Stamos, A., Altsitsiadis, E., & Dewitte, S. (2019). Investigating the effect of childhood socioeconomic background on interpersonal trust: Lower childhood socioeconomic status predicts lower levels of trust. *Personality and Individual Differences, 145*, 19–25.

Statista. (2022). Internetnutzung weltweit. Retrieved from https://de-statista-com.fhooe.idm.oclc.org/statistik/studie/id/6331/dokument/internet-weltweit-statista-dossier/

Statista. (2023a). Statistiken vom Thema Soziale Netzwerke. Retrieved from https://de-statista-com.fhooe.idm.oclc.org/themen/1842/soziale-netzwerke/#topicOverview

Statista. (2023b). Smartphones – Statistiken und Studien. Retrieved from https://de-statista-com.fhooe.idm.oclc.org/themen/581/smartphones/#topicOverview

Stephan, K. E., Penny, W. D., Moran, R. J., Ouden, H. E. M. D., Daunizeau, J., & Friston, K. J. (2010). Ten simple rules for dynamic causal modeling. *NeuroImage, 49*(4), 3099–3109.

Sturgis, P., Read, S., Hatemi, P. K., Zhu, G., Trull, T., Wright, M. J., & Martin, N. G. (2009). A genetic basis for social trust? *Political Behavior, 32*(2), 205–230.

Swamy, A. V., Knack, S., Lee, Y., & Azfar, O. (2001). Gender and corruption. *Journal of Development Economics*, *64*(1), 25–55.

Swan, J. E., Bowers, M. R., & Richardson, L. D. (1999). Customer trust in the salesperson. *Journal of Business Research*, *44*(2), 93–107.

Sztompka, P. (1998). Trust, distrust and two paradoxes of democracy. *European Journal of Social Theory, 1*(1), 19–32.

Takahashi, T., Ikeda, K., Ishikawa, M., Kitamura, N., Tsukasaki, T., Nakama, D., & Kameda, T. (2005). Interpersonal trust and social stress-induced cortisol elevation. *Neuroreport, 16*, 197–199.

Takahashi, T., Ikeda, K., Ishikawa, M., Tsukasaki, T., Nakama, D., Tanida, S., & Kameda, T. (2004). Social stress-induced cortisol elevation acutely impairs social memory in humans. *Neuroscience Letters, 363*, 125–130.

Tarafdar, M., D'Arcy, J., Turel, O., & Gupta, A. (2015). The dark side of IT: Is overuse of information technology sapping your employees' productivity, innovation, and well-being? *MIT Sloan Management Review, 56*(2), 61–70.

Terrell, F., & Barrett, R. K. (1979). Interpersonal trust among college students as a function of race, sex, and socioeconomic class. *Perceptual and Motor Skills, 48*(3), 1194.

Thanetsunthorn, N., & Wuthisatian, R. (2019). Understanding trust across cultures: an empirical investigation. *Review of International Business and Strategy, 29*, 286–314.

Thom, D. H., & Campbell, B. (1997). Patient-physician Trust: an exploratory study. *The Journal of Family Practice, 44*(2), 169–176.

Todorov, A. (2008). Evaluating faces on trustworthiness. *Annals of the New York Academy of Sciences, 1124*(1), 208–224.

Treaster, J. (1999, August 15). Equal Opportunity in Mexico City: Counting on Women to Be More Honest than Men. *New York Times, Section 4*, p. 3.

Turkheimer, E. (1998). Heritability and biological explanation. *Psychological Review, 105*(4), 782–791.

Tzschentke, T. M. (2001). Pharmacology and behavioral pharmacology of the mesocortical dopamine system. *Progress in Neurobiology, 63*(3), 241–320.

Uslaner, E. M. (2018). *The Oxford handbook of social and political trust*. Oxford University Press.

Uzefovsky, F., Bethlehem, R. A. I., Shamay-Tsoory, S., et al. (2019). The oxytocin receptor gene predicts brain activity during an emotion recognition task in autism. *Molecular Autism, 10*(12), 1–13.

Van Den Akker, O., Van Assen, M. A., Van Vugt, M., & Wicherts, J. M. (2020). Sex Differences in Trust and Trustworthiness: A Meta-analysis of the trust game and the gift-exchange game. *Journal of Economic Psychology, 81*(102329), 1–27.

Van Lange, P. A. M., Vinkhuyzen, A. A. E., & Posthuma, D. (2014). Genetic influences are virtually absent for trust. *PLoS ONE, 9*(4), 1–8.

Verbalis, J. (1999). The brain oxytocin receptor(s?). *Frontiers in Neuroendocrinology, 20*, 146–156.

Wadsworth, H., & Tranel, D. (2021). Trust and Lesion Evidence: Lessons from Neuropsychology on the Neuroanatomical Correlates of Trust. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 464–491). Cambridge University Press.

Weinberg, M. K., & Joseph, D. (2017). If you're happy and you know it: Music engagement and subjective wellbeing. *Psychology of Music, 45*(2), 257–267.

Weintraub, D., Hoops, S., Shea, J. A., Lyons, K. E., Pahwa, R., Driver-Dunckley, E., Adler, C. H., Potenza, M. N., Miyasaki, J., Siderowf, A., Duda, J. E., Hurtig, H. I., Colcher, A., Horn, S., Stern, M. B., & Voon, V. (2009). Validation of the questionnaire for impulsive-compulsive disorders in Parkinson's Disease. *Movement Disorders, 24*(10), 1461–1467.

Winston, J. S., Strange, B. A., O'Doherty, J. P., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5*(3), 277–283.

Wood, R. M., Rilling, J. K., Sanfey, A. G., Bhagwagar, Z., & Rogers, R. D. (2006). Effects of tryptophan depletion on the performance of an iterated prisoner's dilemma game in healthy adults. *Neuropsychopharmacology, 31*, 1075–1084.

Wu, Q., Jin, Z., & Wang, P. (2022). The Relationship Between the Physician-Patient Relationship, Physician Empathy, and Patient Trust. *Journal of General Internal Medicine, 6*, 1388–1393.

Yan, Z., & Kirsch, P. (2021). Trust and Oxytocin: Context-Dependent Exogenous and Endogenous Modulation of Trust. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 315–337). Cambridge University Press.

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial Intelligence is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity. *International Journal of Human-Computer Interaction, 37*(10), 981–990.

Zak, P. J., & Fakhar, A. (2006). Neuroactive hormones and interpersonal trust: international evidence. *Economics and Human Biology, 4*(3), 412–429.

Zak, P. J., & Knack, S. (2001). Trust and growth. *The Economic Journal, 111*(470), 295–321.

Zhang, L., Anjum, M. A., & Wang, Y. (2023). The impact of trust-building mechanisms on purchase intention towards metaverse shopping: The moderating role of age. *International Journal of Human-Computer Interaction*. Advance online publication. [URL]

Zheng, S., Masuda, T., Morikura, M., Noguchi, Y., Ohtsubo, Y., Yamasue, H., & Ishii, K. (2020). Oxytocin receptor gene (OXTR) and childhood adversity influence trust. *Psychoneuroendocrinology*, *121*.

6. Anhang: Originalpublikationen

Im Folgenden finden sich die der vorliegenden Dissertation zugrundeliegenden Originalpublikationen.

Nr.	*Referenz*
1	Riedl, R., Hubert, M., & Kenning, P. (2010). Are There Neural Gender Differences in Online Trust? An fMRI Study on the Perceived Trustworthiness of eBay Offers. *Management Information Systems Quarterly, 34*(2), 397-428.
2	Riedl, R., & Javor, A. (2012). The Biology of Trust: Integrating Evidence from Genetics, Endocrinology and Functional Brain Imaging. *Journal of Neuroscience, Psychology, and Economics, 5*(2), 63-91.
3	Riedl, R., Mohr, P., Kenning, P., Davis, F. D., & Heekeren, H. (2014). Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory. *Journal of Management Information Systems, 30*(4), 83-113.
4	Javor, A., Ransmayr, G., Struhal, W., & Riedl, R. (2016). Parkinson Patients' Initial Trust in Avatars: Theory and Evidence. *PLoS ONE, 11*, 1-21.[28]
5	Riedl, R., Javor, A., Gefen, D., Felten, A., & Reuter, M. (2017). Oxytocin, Trust, and Trustworthiness: The Moderating Role of Music. *Journal of Neuroscience, Psychology, and Economics, 10*(1), 1-8.
6	Riedl, R. (2022b). Is Trust in Artificial Intelligence Systems Related to User Personality? Review of Empirical Evidence and Future Research Directions. *Electronic Markets, 32*, 2021-2051.
7	Riedl, R., Hogeterp, S., & Reuter, M. (2024). Do Patients Prefer a Human Doctor, Artificial Intelligence, or a Blend, and is This Preference Dependent on Medical Discipline? Empirical Evidence and Implications for Medical Practice. *Frontiers in Psychology, 15*, 1-13.

[28] Da der Autor der vorliegenden Dissertationsschrift bei dieser Publikation nicht Erst- bzw. Alleinautor ist, sei an dieser Stelle auf die exakte Darstellung seiner umfangreichen Leistungen direkt im Artikel verwiesen (siehe „Author Contributions" in Javor et al. 2016, S. 15-16).

MIS Quarterly

SPECIAL ISSUE

ARE THERE NEURAL GENDER DIFFERENCES IN ONLINE TRUST? AN FMRI STUDY ON THE PERCEIVED TRUSTWORTHINESS OF EBAY OFFERS[1]

By: **René Riedl**
Department of Business Informatics – Information Engineering
University of Linz
Altenberger Strasse 69
4040 Linz
AUSTRIA
rene.riedl@jku.at

Marco Hubert
Department of Corporate Management and Economics
Zeppelin University
Am Seemooser Horn 20
88045 Friedrichshafen
GERMANY
marco.hubert@zeppelin-university.de

Peter Kenning
Department of Corporate Management and Economics
Zeppelin University
Am Seemooser Horn 20
88045 Friedrichshafen
GERMANY
peter.kenning@zeppelin-university.de

[1]David Gefen was the accepting senior editor for this paper. Dennis Galletta served as the associate editor.

Note: Color versions of Figures 3 through 6 are located in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

Abstract

Research provides increasing evidence that women and men differ in their decisions to trust. However, information systems research does not satisfactorily explain why these gender differences exist. One possible reason is that, surprisingly, theoretical concepts often do not address the most obvious factor that influences human behavior: biology. Given the essential role of biological factors—and specifically those of the brain—in decisions to trust, the biological influences should naturally include those related to gender. As trust considerations in economic decision making have become increasingly complex with the expansion of Internet use, understanding the related biological/brain functions and the involvement of gender provides a range of valuable insights.

To show empirically that online trust is associated with activity changes in certain brain areas, we used functional magnetic resonance imaging (fMRI). In a laboratory experiment, we captured the brain activity of 10 female and 10 male participants simultaneous to decisions on trustworthiness of eBay offers. We found that most of the brain areas that encode trustworthiness differ between women and men. Moreover, we found that women activated more brain areas than did men. These results confirm the empathizing– systemizing theory, which predicts gender differences in neural information processing modes.

In demonstrating that perceived trustworthiness of Internet offers is affected by neurobiology, our study has major implications for both IS research and management. We confirm the value of a category of research heretofore neglected *in IS research and practice, and argue that future IS research investigating human behavior should consider the role of biological factors. In practice, biological factors are a significant consideration for management, marketing, and engineering attempts to influence behavior.*

Keywords: Online trust, trustworthiness, functional magnetic resonance imaging (fMRI), gender, eBay

Introduction

Since the days of ancient civilizations, and continuing as modern societies have emerged, trust has played a key role in the prosperous existence of human society (Gambetta 1988; Luhmann 1979). Trust pervades human society in many domains, ranging from friendship to politics. Most notably, trust is central for successful economic exchange (Zak and Knack 2001). In particular, trust is important for economic prosperity in post-modern societies, where an increased emphasis on technologies heightens a perception of complex and uncertain economic relationships (Resnick and Zeckhauser 2002).

Against this background, it is no surprise that during the last decade an increasing number of technology and information systems studies focusing on trust have been published (e.g., from Hoffman et al. in 1999 to Kim et al. in 2009). Trust has typically been conceptualized as an attitude, belief, intention, or behavior in the existing IS literature (McKnight and Chervany 2001). As a consequence, the existing body of knowledge is primarily focused on the behavioral level of analysis.

In the present article, we borrow concepts, methods, and theories from biology (more precisely, from neuroimaging and neurobiology) in order to gain new insights into trust. In this way, our study adds a significant and expanded view to the question of the nature of trust in IS research. Our approach is in line with other research in human trusting behavior that addresses the role of genes (e.g., Cesarini et al. 2008), hormones (e.g., Kosfeld et al. 2005; Zak et al. 2005), and neurology (e.g., Baumgartner et al. 2008; King-Casas et al. 2005). The ultimate goal of these biological research streams is to explain a maximum of the variance of human trusting behavior (Fehr and Camerer 2007; Kenning and Plassmann 2005).

The present study examines gender differences in online trust. In particular, we investigate the following research question: *Are there neural gender differences in online trust?* This question explicitly addresses (1) the different brain areas that women and men activate in online trust situations and (2) whether women and men utilize a varying number of brain areas.

This research topic is important for a number of reasons. First, IS research has found notable gender effects in IT perception and usage (Gefen and Ridings 2005; Gefen and Straub 1997; Venkatesh and Morris 2000). More recent empirical evidence shows that these differences apply to online trust as well (Awad and Ragowsky 2008). Therefore, understanding how these differences relate to brain activity could shed light on why and how these gender differences occur.

A second foundation for this study focuses on gender as a variable that can be easily integrated into the management of information and communication systems. Hence, if our study reveals neurobiological gender differences in online trust, this knowledge could yield fruitful insights for IT managers, engineers, and marketers. If, for example, the female and male brain process online offers differently (resulting in varying trust levels), and with the knowledge that most Internet platforms identify the gender of registered users, then offers could be tailored to gender-specific requirements in order to increase perceived trustworthiness. A case in point would be developing Internet sites with real-time interface adaptation capacities (e.g., with respect to such factors as content, information presentation, or color) based on the user's gender (Simon 2001; Stenstrom et al. 2008).

A third rationale for our research topic, posing a particularly promising argument, is the key role of gender as a variable in neuroscience research (Bell et al. 2006; Cahill 2006; Cosgrove et al. 2007; Haier et al. 2005). The success of our study, therefore, could indicate an opportunity for neuroscientific findings on human behavior to be integrated into IS research.

Perhaps the strongest reason for choosing this research topic is to use a neurobiological focus to add a new theoretical explanation for the existence of gender-specific differences in IT-related behavior to the current body of literature (e.g., Gefen and Ridings 2005; Gefen and Straub 1997; Venkatesh and Morris 2000). Altogether, our approach is in line with that of the limited number of published NeuroIS papers (Dimoka and Davis 2008; Dimoka et al. 2007), and with that of Gefen et al. (2008), who state that "the trust literature in online environments can substantially benefit from the neuroscience literature and functional neuroimaging studies" (p. 283). In this sense, our study aims to build a bridge between the predominantly independent research fields of IS and neurobiology.

Against this background, the remainder of this paper is structured as follows. In the literature review, we begin with a brief overview of gender differences in general trust behavior. Next, we outline studies focused on general gender differences in the IT realm, before discussing important (though more limited) literature that is focused specifically on gender differences in online trust and related behaviors (e.g., risk perception). In addition to the behavioral literature, we present neuroscience knowledge on gender differences in brain anatomy and functioning. The hypotheses that follow are based on both behavioral and neuroscience work, and are the core concepts that we investigated through functional brain imaging. The next section describes the research methodology in detail. An outline of the research results precedes a more thorough analysis of the results, wherein we connect the IS literature with the neuroscience literature. Finally, we summarize our findings, outline implications, present limitations, and provide concluding comments.

Literature Review and Hypotheses

The expansion of the role of the Internet has led to a recent identification of trust in online environments as a central IS research topic (Steininger et al. 2009). Several research streams have been developing, particularly investigation into (1) trust in virtual teams (e.g., Gallivan 2001; Jarvenpaa et al. 2004; Paul and McDaniel 2004; Piccoli and Ives 2003); (2) trust in interorganizational collaboration (e.g., Allen et al. 2000; Gallivan and Depledge 2003; Nicolaou and McKnight 2006; Welty and Becerra-Fernandez 2001); (3) trust in e-government (e.g., Carter and Belanger 2005; Grimsley and Meehan 2007; Horst et al. 2007; Teo et al. 2008); (4) trust in IT artifacts such as recommendation agents or avatars (e.g., Komiak and Benbasat 2006; Qiu and Benbasat 2005; Vance et al. 2008; Wang and Benbasat 2008); (5) trust in virtual worlds and virtual communities (e.g., Fogel and Nehmad 2009; Johnson and Kaye 2009; Junglas et al. 2007; Oxendine et al. 2003); and, finally, (6) trust in e-commerce (e.g., Ba and Pavlou 2002; Gefen et al. 2003; McKnight et al. 2002a; Pavlou and Gefen 2004).

However, given the focus of the present study, we limit our literature review to gender differences in general trust behavior, general gender differences in the IT realm, and gender differences in online trust and related behaviors (e.g., online shopping risk perceptions). In addition to these gender differences on the behavioral level of analysis, we also discuss gender differences in brain anatomy and functioning. Drawing upon both behavioral and neurological gender research, we deduce two main hypotheses (Hypotheses 1 and 2) and seven sub-hypotheses (Hypotheses 1.1 through 1.7).

Gender Differences in General Trust Behavior

With respect to the literature on trusting others, both survey results and experimental findings show that, in general, *men trust more than do women*. Several surveys found evidence that women are less likely to believe that "most people can be trusted" (Alesina and La Ferrara 2002; Glaeser et al. 2000; Terrell and Barrett 1979). A number of investigations in experimental economics support those findings, concluding that men are more trusting than women. The determination is based on results showing that men sent more money to their opponents in economic games, thereby being more vulnerable to the action of the other party (e.g., Buchan et al. 2008; Snijders and Keren 1999).

In the literature investigating the concept of trusting others in view of gender, empirical evidence shows that, in general, *women are more trusted than are men*. A study of psychological attitudes, using Rotter's (1967) interpersonal trust scale, revealed greater trust toward women than toward men (Wright and Sharp 1979). Moreover, in a study focused on economic behavior, male clients were believed to be less trustworthy than female clients (Shaub 1996).

With respect to a determination that one gender is more trustworthy than the other, both institutional arrangements and scientific evidence indicate that, in general, *women are more trustworthy than men*. *The New York Times* reported that teams of female traffic police were formed in Mexico City, in the belief that female officers are less likely to take bribes. When asked about the reason for this unorthodox arrangement, the police chief responded simply, "I trust them," illustrating the view that women are more trustworthy than men (Treaster 1999). Scientific research corroborates both the notion that women are less likely to take bribes, and the belief that corruption is negatively correlated with the rate of female participation in the public domain (Dollar et al. 2001; Swamy et al. 2001). In line with these results, experimental evidence from economic games shows that women reward trust through financial reciprocation more often than do men, hence, being more trustworthy (Buchan et al. 2008; Croson and Buchan 1999; Snijders and Keren 1999). Table 1 summarizes the studies that investigate gender differences in general trust behavior.

General Gender Differences in the IT Realm

In the IT realm, a substantial number of empirical studies document gender differences in, for example, the use of computers and the Internet, attitudes toward computers, levels of computer anxiety, computer skills, perceptions of communi-

Table 1. Gender Differences (Summary of Literature Review)	
Gender differences in general trust behavior	***Sources***
Men trust more than do women.	Alesina and La Ferrara (2002), Buchan et al. (2008), Glaeser et al. (2000), Snijders and Keren (1999), Terrell and Barrett (1979)
Women are more trusted than are men.	Shaub (1996), Wright and Sharp (1979)
Women are more trustworthy than men.	Buchan et al. (2008), Croson and Buchan (1999), Dollar et al. (2001), Snijders and Keren (1999), Swamy et al. (2001), Treaster (1999)
General gender differences in the IT realm	***Sources***
Women and men differ in their perceptions of communication technologies such as virtual communities and e-mail.	Gefen and Ridings (2005), Gefen and Straub (1997)
Computer usage decisions of women are more strongly influenced by a system's ease of use, whereas men's decisions are more strongly influenced by usefulness.	Venkatesh and Morris (2000)
Men use computers and the Internet more often than women.	Seybert (2007)
Women feel less competent and comfortable with the computer and the Internet than men.	Schumacher and Morahan-Martin (2001)
Women exhibit more negative attitudes toward computers and the Internet than men.	Broos (2005)
Women exhibit higher levels of computer anxiety than men.	Broos (2005), Jackson et al. (2001)
More men than women study computer science, are employed in computing jobs, and work in IT-related academic disciplines such as computer science.	Frenkel (1990), Seybert (2007), Van Welsum and Montagnier (2007)
More men than women have basic computer skills.	Seybert (2007), Van Welsum and Montagnier (2007)
Women acquire their computer skills differently than men, that is, they learn from colleagues and friends or from learning-by-doing rather than in formal courses.	Van Welsum and Montagnier (2007)
Women use computers and the Internet for different reasons than men; for example, they search online for health information, and send and receive e-mails, rather than playing games or downloading software.	Jackson et al. (2001), Van Welsum and Montagnier (2007), Weiser (2000)
Women and men differ in their Internet acceptance and usage patterns.	Sánchez-Franco (2006)
Significant gender differences do exist in perceptions of website design and website satisfaction.	Cyr and Bonanni (2005)
Gender differences in online trust and related behaviors	***Sources***
Men are more likely to intend to use the Web for making purchases than women. Men rate the trustworthiness of Web shopping higher, and its complexity lower, than do women.	Van Slyke et al. (2002)
Women are more likely to cue into the more detailed aspects of an online shopping website than men, relying on the extent to which those cues affect them emotionally and drive them.	Rodgers and Harris (2003)
Women, in contrast to men, perceive a higher level of risk in online shopping. Having a website recommended by a friend leads to both a greater reduction in perceived risk and a stronger increase in willingness to buy online for women than for men.	Garbarino and Strahilevitz (2004)
Women perceive a greater risk (i.e., privacy concerns) than do men in online shopping.	Sheehan (1999)
Men are more convenience-oriented and less motivated by social interaction than women in online shopping.	Swaminathan et al. (1999)
Whereas socio-psychological and emotional factors are more important for women than functional concerns in conventional buying, this order reverses when women buy online. Moreover, the functional factors, which have always played an important role for men, are even amplified in the shift from conventional buying to online shopping.	Dittmar et al. (2004)
Men exhibit more positive beliefs about and more positive attitudes toward Web advertising. Men are more likely than women to purchase from the Web, although women are more likely to surf the Web for shopping reasons.	Wolin and Korgaonkar (2003)
The influence of trust on commitment and of commitment on loyalty is stronger for women than men, while the effects of satisfaction on commitment and of trust on loyalty are stronger for men than for women.	Sánchez-Franco et al. (2009)
The effect of trust on intention to shop online is stronger for women than for men. Women value the responsive participation of other consumers to the content they have posted in online word-of mouth systems, whereas men value their ability to post content. (In online word-of-mouth systems, consumers can rate products offered for sale.)	Awad and Ragowsky (2008)

cation technologies such as e-mail, and perceptions of website design. Table 1 summarizes studies from various disciplines (e.g., psychology, marketing, IS) that investigate general gender differences in the IT realm.

In the IS discipline, important milestones in gender research have been set by Gefen and Straub (1997), as well as by Gefen and Ridings (2005); both papers investigate gender differences in the perception and use of communication technologies (e-mail and virtual communities). Major findings revealed by these studies are that women and men differ in their perceptions of e-mail, but not in their actual use of e-mail (Gefen and Straub 1997), and that in virtual communities men communicate to establish superior social standing, while women communicate with an undertone of rapport, compassion, and empathy (Gefen and Ridings 2005). These differences affect the respective perceptions of community quality for men and women. Another influential gender study, conducted by Venkatesh and Morris (2000), investigated differences between men and women in the adoption and use of technology. The study's primary finding is that computer usage decisions of women are more strongly influenced by a system's ease of use, whereas men's decisions are more strongly influenced by usefulness. Building from this research, the following section focuses on gender differences in online trust and related behaviors (e.g., online shopping risk perceptions).

Gender Differences in Online Trust and Related Behaviors

Extensive investigation revealed a very limited number of published studies dealing with gender differences in online trust and related behaviors, indicating an area with significant potential for research. Most notably, the articles we discuss here can be considered as a body of highly fragmented work published in varied fields of research (primarily in IS, marketing, and psychology) and are, thereby, hardly cumulative in nature. Nevertheless, these articles constitute a valuable empirical base for the present study, particularly because the findings substantiate significant gender differences that are important to the focus of our paper.

One study (Van Slyke et al. 2002) investigating gender differences in perceptions of Web-based shopping determined that men, in contrast to women, are more likely to use the Web with intent to make purchases. Their study also found that men's perceptions of the characteristics of Web shopping are more favorable than are women's perceptions. Specifically, men rated the trustworthiness of Web shopping higher, and gave lower ratings to its complexity, than did women in the study.

Another study (Rodgers and Harris 2003) investigated the reasons why women are less satisfied than men with the online shopping experience. The findings indicate that emotion, trust, and convenience are predictors of both women's dissatisfaction with online shopping and men's satisfaction with shopping online, and are predictors of the actual shopping behaviors of men and women. Moreover, the study determined that women, in contrast to men, are more likely to cue into the more detailed aspects of a Web site, relying on the extent to which those cues affect them emotionally and drive them. Conversely, the study found that men are more likely to consider a shopping site in terms of a general attitude than are women, indicating that affective cues do not drive male behavior as much as they do the behavior of women.

In a study investigating the perceived risk of buying online, Garbarino and Strahilevitz (2004) found that, in contrast to men, women perceive a significantly higher likelihood of negative outcomes in online shopping (e.g., credit card misuse or shipping problems). Moreover, the study revealed that women also perceive the consequences of negative events to be more severe. Given the fact that perceived risk is a combination of the perception of the likelihood that something will go wrong and the perception of the seriousness of the consequences if it does (Bettman 1973), the Garbarino-Strahilevitz study concludes that women, in general, perceive a higher level of risk in online purchasing than do men, thereby corroborating findings of a similar study (Sheehan 1999) on gender differences in Internet shopping risk perceptions. Moreover, Garbarino and Strahilevitz's study found that, for women, having a site recommended by a friend leads to both a greater reduction in perceived risk and a stronger increase in willingness to buy online than it does for men, which indicates that trust-building mechanisms (in their study, a site recommendation by a friend) affect trust perceptions of women and men differently.

Gender differences in conventional shopping and online shopping have also been the focus of research, and reveal that shopping generally plays a much more emotionally encompassing role for women than for men (Campbell 2000; Dittmar and Drury 2000). That is, women have highly positive attitudes toward shopping, associating it with a leisure frame, whereas men tend to have more negative attitudes toward buying, viewing it as work that should be accomplished with minimum input of time and effort. Women, therefore, tend to focus on the often enjoyable *process* of buying, whereas men primarily focus on the *outcome* of obtaining the goods (Dittmar et al. 2004). In other words, men are more functional in their buying attitudes than women who, in turn, are more inclined to emphasize socio-psychological and emotional concerns (Dittmar et al. 1996).[2]

With respect to online shopping, one study (Swaminathan et al. 1999) found that male buyers were more convenience-oriented and less motivated by social interaction than were female buyers, thereby corroborating existing evidence from studies of conventional buying patterns (Dittmar and Drury 2000). In contrast, another study (Dittmar et al. 2004) comparing conventional buying with online shopping found a reversal of buying attitudes. Although in conventional buying socio-psychological and emotional factors are more important than functional concerns for women, the order of significance reverses when women buy online. For men, the functional factors, which typically play an important role, are amplified in the shift from conventional buying to online shopping. In online shopping, women's attitudes become more similar to those of men; that is, both genders consider functional concerns relevant, even though men are comparatively more oriented toward functional factors than women. Given this empirical finding, the shift from conventional buying to Internet shopping implies a much more significant attitudinal change for women, with the possible result that female Internet users experience an increased level of discomfort and arousal. Such an observation might help to explain why women exhibit lower levels of trust in Internet shopping than men (Dittmar et al. 2004; Sheehan 1999; Van Slyke et al. 2002), because arousal and trustworthiness are interrelated (Ravaja et al. 2004). Taken together, the findings of Dittmar et al. (2004) challenge those of Swaminathan et al. (1999). Additional empirical evidence is needed, therefore, to clarify whether socio-psychological and emotional concerns remain most important for women in online shopping, or whether functional factors become the more relevant category.

Another study (Wolin and Korgaonkar 2003) investigated gender differences in consumer beliefs about Web advertising. The results of the study indicate that men exhibit more positive beliefs about and attitudes toward Web advertising than women. Moreover, the study reveals that men are more likely to purchase from the Web than are women, although women are more likely to surf the Web for shopping reasons (rather than to surf the Web for entertainment reasons as do men).

[2]Important functional factors are, for example, possibilities to compare prices as well as opportunities to save time while buying goods. In contrast, important socio-psychological and emotional factors are possibilities to experience the shopping process as a leisure activity as well as the opportunity to see and touch goods before purchase (Dittmar et al. 2004).

A further study (Sánchez-Franco et al. 2009) investigated online customer loyalty toward an Internet service provider, based on commitment-trust theory (Morgan and Hunt 1994). The study found that the influence of trust on commitment, and that of commitment on loyalty, was significantly stronger for women than men, while the effects of satisfaction on commitment and of trust on loyalty were significantly stronger for men. Finally, another study (Awad and Ragowsky 2008) investigated the effect of gender on the relationship between online word-of-mouth quality and online trust. (In online word-of-mouth systems, consumers can rate products offered for sale.) The results of this study reveal that the effect of trust on intention to buy online is stronger for women than for men. Moreover, the study found that men value their ability to post content online, whereas women value the responsive participation of other consumers to the content they have posted.

Given the combined research on gender differences in general trust behavior, general gender differences in the IT realm, and the gender differences in online trust and related behaviors, evidence supporting substantial behavioral differences between women and men is available. These gender differences on the behavioral level should be associated with differences on the biological level, because all human behavior is—at least partly—determined by biological factors, in particular those related to the brain (Turkheimer 1998). Moreover, recent brain research has revealed notable gender differences with respect to brain anatomy and functioning (Cahill 2006; Cosgrove et al. 2007; Haier et al. 2005). Considering this existing empirical knowledge, there is reason to state the following prediction:

> *Hypothesis 1: Decision making on both trustworthy and untrustworthy Internet offers activates* ***different brain areas*** *in men and women.*

Gender Differences in Brain Anatomy and Functioning

Investigations of neurobiological gender differences have a long history. For example, it is a well-established fact since the 19th century that the male brain is larger than the female brain (Broca 1861; Darwin 1871). Using sophisticated techniques such as magnetic resonance imaging, and correcting for body size, more recent evidence has confirmed this finding (Rushton 1992; Rushton and Ankney 1996), indicating that the cranial capacity of men averages 1,442 cm^3 and that of women averages 1,332 cm^3.

In the recent past, a number of additional gender-specific differences were identified. As an example, the corpus callosum, the primary means of communication between the two cerebral hemispheres, is, on average, larger in women. This condition indicates that brain functions requiring connectivity between the two hemispheres (e.g., language) can be better performed by women (Halpern et al. 2007). Other brain regions which are, on average, larger in women include the caudate nucleus and hippocampus. In contrast, men usually have larger amygdalae and hypothalami (Cosgrove et al. 2007). Gender-specific size differences of certain brain regions may affect performance in decision making, memory, and learning tasks (Cahill 2006), thereby being of particular interest for NeuroIS researchers.

With respect to the analysis of gender differences in brain functioning, the *empathizing–systemizing theory* (Baron-Cohen et al. 2005) has gained considerable attention during the past few years. Empathizing is the ability to predict and respond to the behavior of agents (usually people) by inferring the mental states of the agents and responding with an appropriate emotion. Systemizing is the ability to predict and to respond to the behavior of nonagentive deterministic systems by analyzing input–operation–output relations and deducing the rules that govern such systems. The empathizing–systemizing theory predicts that, at a population level, women are stronger empathizers and men are stronger systemizers.

Drawing upon this theory, a number of gender studies in neuroscience (e.g., Mak et al. 2009; McClure et al. 2004; Schulte-Rüther et al. 2008) found that women more often activate emotion-related brain regions in social interaction tasks (in particular *limbic structures* such as the anterior cingulate cortex, thalamus, and hippocampus), whereas men more often activate brain regions associated with cognitive information processing (in particular *prefrontal structures* such as the dorsolateral prefrontal cortex and ventromedial prefrontal cortex). Therefore, in a trustworthiness evaluation task such as was used in the present study, we expect to see that women recruit more limbic structures than men, who in turn should recruit primarily prefrontal structures.

To the best of our knowledge, no published neuroscience study on trust has focused on gender differences. However, the results of the studies available (Baumgartner et al. 2008; Delgado et al. 2005; King-Casas et al. 2005; Krueger et al. 2007; Winston et al. 2002) show that both the limbic and prefrontal systems play a crucial role in trust situations. Moreover, two additional brain regions, the *striatum* and *insular cortex*, were shown to be important in trust situations (e.g., Baumgartner et al. 2008). In the following sections, we discuss brain regions that have been shown to encode human trust, and demonstrate connections between this literature and existing knowledge of gender-specific differences in brain functioning. Drawing upon this evidence, we deduce seven sub-hypotheses about gender-specific brain activation differences in online trust.

Anterior Cingulate Cortex (ACC)

The ACC has been studied intensively in neuroscience since the late 1990s, with demonstrated evidence that its activation is associated with a number of specific cognitive functions. Carter et al. (1998) state that the ACC detects conditions under which errors are likely to occur. Similarly, Botvinick et al. (1999) argue in favor of a conflict-monitoring account of the ACC function. In another investigation, Botvinick et al. (2004) reviewed 80 papers on ACC functions and conclude that "the ACC might monitor conflict as an index of task difficulty (and/or the mental effort it demands), entering this into cost–benefit analysis underlying action or strategy selection" (p. 545).

With respect to online trust, the ACC may have a critical role in processing information displayed on an Internet shopping site, because conflict in information processing and the associated mental effort could result from a comparison of the potential benefits and risks associated with an Internet offer. In line with this view, recent research (Seo and Lee 2007) has found that neurons in the ACC often encode signals related to expected or actual rewards (benefits), whereas another study (Milad et al. 2007) found that cortical thickness within the ACC is positively correlated with skin conductance response, which is a measure for fear expression (risks).

Anatomically, the ACC can be subdivided into dorsal and ventral parts. Brain damage studies and functional imaging studies both indicate separate roles for the dorsal (BA 32, Brodmann area) and ventral (BA 24) part of the ACC (Bush et al. 2000). The dorsal ACC is more involved in a cognitive distributed cortical network, including lateral prefrontal, parietal, premotor, and supplementary motor areas (Bush et al. 2000; Devinsky et al. 1995). The ventral ACC, in contrast, has strong connections to limbic and paralimbic brain regions such as the amygdala and the orbitofrontal cortex (Devinsky et al. 1995). Hence, the ventral ACC, especially, is involved in assessing the salience of emotional stimuli and in regulating emotional responses (Bush et al. 2000) and is, thereby, associated with an affective network of brain structures.

According to the empathizing–systemizing theory, men are stronger systemizers and women are stronger empathizers. This implies that men use more cognitive brain structures during information processing, while women use more affective structures. Drawing upon this argument, we predict

Hypothesis 1.1: Decision making on both trustworthy and untrustworthy Internet offers activates the dorsal ACC (BA 32) in men and the ventral ACC (BA 24) in women.

Thalamus

Research has indicated a role for the thalamus in attentiveness and in the capacity to perform tasks of rapid information processing. Van der Werf et al. (2001), for example, found that the size of the thalamus predicted performance on tasks that require cognitive speed. Further research has shown that the thalamus is also associated with reward processing. It was found that single thalamic neurons can code for the acquired significance of sensory stimuli and can predict future reward value (Komura et al. 2001). Because trustworthy rather than untrustworthy Internet offers are likely to trigger activation in reward regions, there is reason to believe that thalamus activation can be found during the processing of trustworthy Internet offers. With respect to gender differences in thalamus activation, a meta-analysis has determined that in emotional situations the thalamus is activated in women more often than in men (Wager et al. 2003). Considering this evidence, we predict

Hypothesis 1.2: Decision making on trustworthy Internet offers activates the thalamus significantly more in women than in men.

Hippocampus

The hippocampus is well known for its role in processing information to be stored in long-term memory (e.g., Alkire et al. 1998). In addition to this memory function, recent animal experiments show that this brain structure is also associated with the processing of fearful and/or anxiogenic stimuli (Bannerman et al. 2004; McHugh et al. 2004). Although caution is necessary when using results of animal studies for the interpretation of human brain activation and behavior, one study in the trust domain (Winston et al. 2002) found that processing untrustworthy faces may result in parahippocampal activation, thereby creating support for application of the animal experiment findings. Altogether, there is reason to believe that untrustworthy Internet offers may activate the hippocampus and parahippocampus areas, respectively, due to their uncertainty and risk-signaling functions. Anatomically, it is a well-established fact that the female hippocampus is, on average, larger than that of the male (Cosgrove et al. 2007). This neurological gender difference could contribute to women's heightened perception of uncertainty and risk (e.g., Byrnes et al. 1999). Drawing upon the presented evidence, we therefore predict

Hypothesis 1.3: Decision making on untrustworthy Internet offers activates the hippocampus significantly more in women than in men.

Dorsolateral Prefrontal Cortex (DLPFC)

Research revealed brain activation in the DLPFC in trust situations (Krueger et al. 2007). Goal-directed behavior, which plays a crucial role in trust situations (Fehr and Camerer 2007), requires information regarding whether or not implemented actions were successful in obtaining outcomes. For example, if a person makes a decision to trust another person, expected rewards may or may not be realized. The DLPFC was found to play a major role in coupling the information of rewards to actions, thereby controlling behavior and optimizing decision making (Heekeren et al. 2006; Lee and Seo 2007; Van't Wout et al. 2005).

In contrast to women, men usually process information in a cognitive rather than affective manner (Cosgrove et al. 2007; Meyers-Levy 1994; Wager et al. 2003). Coupling information about rewards to actions and, thereby, controlling behavior and optimizing decision making, is a cognitive rather than affective process, because the former implies the deliberate processing of several information cues and not simply the unconscious processing of a few cues, as does the latter. Drawing upon this line of reasoning, we predict

Hypothesis 1.4: Decision making on both trustworthy and untrustworthy Internet offers activates the DLPFC significantly more in men than in women.

Ventromedial Prefrontal Cortex (VMPFC)

In the 1990s, a number of clinical case reports revealed a central role for the VMPFC in human decision making (Bechara et al. 1996, 1997, 1999; Damasio 1996). These reports show that, similar to the patterns of young children (Crone and Molen 2004), patients with VMPFC damage are insensitive to future consequences, positive or negative, and are primarily guided by immediate prospects. Notably, this so-called "myopia for the future" persists even in the face of severe negative consequences such as financial loss or declining future reward (Bechara et al. 2000). Given this finding, the important role of the VMPFC in shopping becomes evident (Deppe et al. 2005). In contrast to ordinary shopping, online shopping is often more risky because of the anonymity afforded by the Internet. Consequently, the deliberate evaluation and anticipation of the consequences of an online transaction (both short- and long-term) are crucial in order to avoid losses as a result of deception.

In addition to the findings of clinical studies, activation in the VMPFC may represent information about value (Gläscher et al. 2009; O'Doherty 2004; Paulus and Frank 2003). Discussing the role of the VMPFC in decision making, Fellows and Farah (2007) stress that the essence of decision making is the weighting of the value of options, and value is not a simple and fixed feature of a stimulus. Rather, it is relative and context-dependent. One such highly important context variable in online environments is uncertainty, which has been shown to be associated with activity changes in the VMPFC (Fellows and Farah 2007). Hence, VMPFC activation may be associated with the processing of untrustworthy Internet offers.

With respect to gender differences in VMPFC activation, one study (Shirao et al. 2005) found that in an emotional decision task, VMPFC was activated in men, but not in women. Shirao et al. (2005) attribute this result to gender differences in neural information processing:

> The gender differences detected in our study may demonstrate differences of cognitive pattern in men and women. Our results suggest the possibility that men processed the emotional decision task including words concerning body image more cognitively rather than emotionally, and activation in the medial prefrontal cortex was prominent; on the other hand, women processed this task more emotionally rather than cognitively, and the medial prefrontal cortex did not exhibit any significant activation (p. 51).

Considering this evidence, which obviously provides strong support for the empathizing–systemizing theory, we predict

> *Hypothesis 1.5: Decision making on untrustworthy Internet offers activates the VMPFC significantly more in men than in women.*

Striatum

It is well known that activation in the human striatum is associated with value and reward, as well as with anticipation of the value and reward (O'Doherty et al. 2004; Tricomi et al. 2004). Previous fMRI studies in neuroeconomics have shown that activations in both the caudate nucleus and putamen—which are major parts of the striatum—are associated with trusting intentions and social cooperation (Baumgartner et al. 2008; Delgado et al. 2005; King-Casas et al. 2005). This result is in line with the striatum's value/reward function, because a major goal of trusting and cooperating with another person is to realize value and reward, respectively.

Dopamine, an important neurotransmitter, is closely associated with the striatum. Dopamine has been found to be correlated with pleasure, value, and reward in the brain, providing positive feelings and reinforcement to motivate proactive behavior to perform certain activities. Furthermore, dopamine is released by naturally rewarding experiences and also by a specific part in the striatum, the nucleus accumbens (e.g., Schultz et al. 1997). With respect to gender differences, it is a well-established fact that the dopaminergic function is enhanced in women, as compared to men (Cosgrove et al. 2007). Moreover, research found that the female caudate nucleus is, on average, larger than that of the male (Cosgrove et al. 2007). Considering these existing research findings, we predict

> *Hypothesis 1.6: Decision making on trustworthy Internet offers activates the striatum significantly more in women than in men.*

Insular Cortex

Research shows that insular activation correlates with (1) the perception of faces that express a feeling of disgust (Phillips et al. 1997, 1998); (2) the anticipation of emotionally aversive visual stimuli such as spiders and snakes (Simmons et al. 2004); and (3) the anticipation of physical pain (Ploghaus et al. 1999). With respect to buying behavior, Knutson et al. (2007) found that excessive prices may activate the insular in purchase situations. Additionally, several studies have found that the insular is activated in uncertain decision situations (Chritchley et al. 2001; Ernst et al. 2002; Paulus et al. 2003), in risky decision situations (Clark et al. 2008; Kuhnen and Knutson 2005; Preuschoff et al. 2006, 2008), and in ambiguous decision situations (Huettel et al. 2006; Krain et al. 2006). Because women, in contrast to men, perceive higher levels of uncertainty and risk (e.g., Byrnes et al. 1999), we predict

> *Hypothesis 1.7: Decision making on untrustworthy Internet offers activates the insular cortex significantly more in women than in men.*

Gender Differences in the Number of Activated Brain Areas

So far, we have discussed brain regions that were shown to play a crucial role in trust situations. We connected these regions to neuroscience knowledge of both anatomical and information-processing gender differences. Drawing upon this evidence, we conceptualized our predictions about gender-related activation differences in specific brain areas in seven sub-hypotheses. In four of these hypotheses, we predict that the respective brain region is activated significantly more in women than in men (thalamus, hippocampus, striatum, and insular cortex). In two hypotheses, we predict that the region is activated significantly more in men than in women (DLPFC and VMPFC). Finally, in one hypothesis we predict that specific parts of the brain region are activated differently in men and women (ACC). Altogether, we have reason to anticipate that women will recruit more brain areas than men in our trustworthiness evaluation task.

Moreover, women's greater inclination toward risk perception is likely to reveal a gender difference in the number of brain areas activated in trust situations. As discussed, women usually perceive a higher level of risk in online purchasing than do men (e.g., Garbarino and Strahilevitz 2004). A possible strategy for female consumers to cope with this higher level of risk perception is to search for and consider more information, because information can result in a higher level of confidence in a decision (Hall et al. 2007). The deliberate information search and consideration could, however, be reflected in a more sophisticated processing within the human brain.

This theorizing is supported by the *selectivity model* (Meyers-Levy and Maheswaran 1991; Meyers-Levy and Sternthal 1991; Simon 2001), which evaluates gender-specific differences in information processing. This model indicates that men often do not engage in comprehensive processing of all available information, but instead are selective; that is, they consider and rely only on subsets of highly available cues, thereby activating fewer brain areas in contrast to women. Women, comparatively, tend to use more comprehensive information processing strategies, thereby considering a larger number of available cues and engaging in more effortful and particularized analysis of all available information (Simon 2001). In line with this argument, one study (Rodgers and Harris 2003) in the online shopping domain found that men, in contrast to women, typically evaluate a shopping site in terms of a general attitude. Conversely, women typically cue into the more detailed aspects of a Web site, relying on the extent to which those cues affect them emotionally and drive them. This, obviously, accompanies the activation of a greater number of brain regions, that is, a more sophisticated pattern of brain areas. Drawing upon this evidence, we predict

> *Hypothesis 2: Decision making on both trustworthy and untrustworthy Internet offers activates **more brain areas** in women than in men.*

To the best of our knowledge, no available study investigates the hypotheses presented here. Against this background, the next section describes the research methodology through which we were able to demonstrate that for women the decision to trust an Internet offer can trigger activation in more and in different brain areas than is the case for men, which in turn is reflected in different trustworthiness ratings.

Research Methodology

Stimulus Development

Previous research has demonstrated that trust-assuring arguments, which are formulated on the basis of Toulmin's (1958) model of argumentation, affect consumers' trusting beliefs (Kim and Benbasat 2006). Accordingly, we used Toulmin's model as a basis to create product descriptions, thereby influencing the *perceived trustworthiness of Internet offers*, which we then used as the stimulus material in our fMRI experiment. In line with previous IS research (e.g., Gefen et al. 2003; Kim and Benbasat 2006; McKnight et al. 2002a), we define perceived trustworthiness of an Internet offer as a consumer's trusting belief, which refers to the aggregation of an offer's positive characteristics.

Toulmin's model proposes a layout containing six interrelated components for analyzing arguments (Figure 1): CLAIM (an assertion or conclusion put forward for general acceptance, always having a potentially controversial nature), DATA (statements specifying the particular facts or previously established beliefs about a situation, as a basis on which a claim is made), WARRANT (a statement that justifies or certifies the reasonableness of moving from data to a claim), BACKING (the general body of information or experience that assures the trustworthiness of a warrant), QUALIFIER (articulation of the degree of certainty associated with a claim), and REBUTTAL (an extraordinary or exceptional circumstance that can defeat the warranted claim).

Figure 1 illustrates an example of the interrelationships among the six components. The illustration draws upon a product description text of a fictitious eBay seller. In our experiment, the product for sale was a USB flash drive. We used the text blocks in Figure 1 to create the product descriptions. Considering the various levels of arguments illustrated in Figure 1, several levels of trust-assuring description texts become evident: (1) NO TRUST-ASSURING ARGUMENT (that is, a seller provides no description text), (2) CLAIM ONLY, (3) CLAIM + DATA, (4) CLAIM + DATA + BACKING, and (5) CLAIM + DATA + BACKING + REBUTTAL. Additionally, QUALIFIERS (certainly, presumably, probably, and possibly) may affect the trustworthiness of a product description text.[3]

[3]Note that we did not use the WARRANT to manipulate the product description text, because it is usually not stated explicitly in argumentations; rather, it is substantiated by the BACKING (Toulmin 1958).

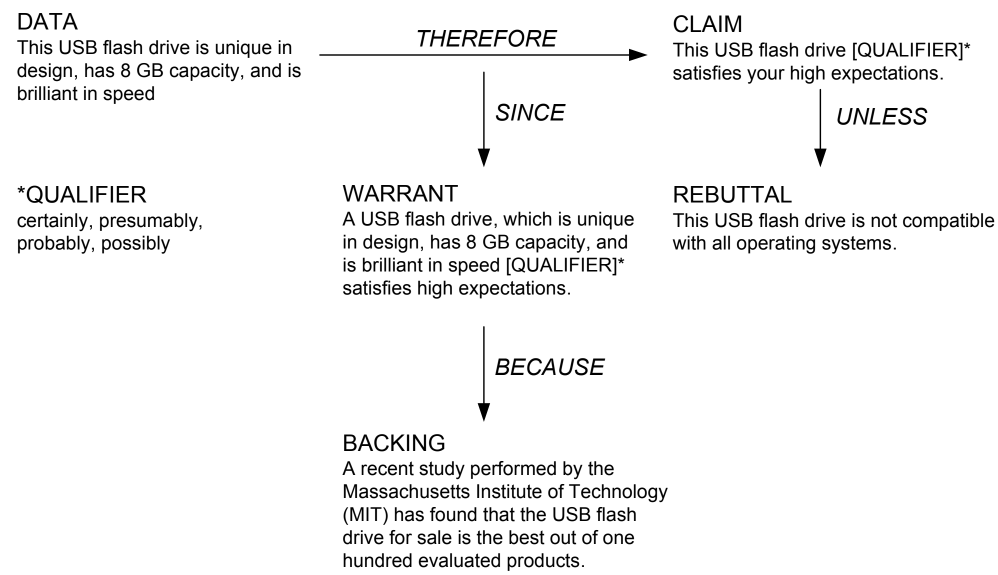


Figure 1. Product Description Text and Toulmin's Model of Argumentation

To enhance the external validity of our experiment, we embedded the elements of Toulmin's model into a realistic eBay site containing the following characteristics (see Figure 2): eBay logo, product name (New USB Flash Drive), picture of the product, selling mode (Buy It Now, indicating our decision to use eBay's online shopping feature rather than a bidding auction), price (EUR 30.00), seller's name (usb-shop-123), seller's experience level (55, with a blue star indicating that 50 to 99 feedback ratings have been posted), feedback (100% positive), duration and location of membership (since October 6, 2004, in Germany), and, finally, the product description. Figure 2 shows a stimulus example with the maximum amount of text allowed in the product description field, that is, 69 words: CLAIM + QUALIFIER (10 w.) + DATA (17 w.) + BACKING (29 w.) + REBUTTAL (13 w.).[4] It is important to note that the seller's name (i.e., usb-shop-123) was designed to eliminate the possibility of any inference concerning gender and subsequent potential for bias, based on the knowledge that women are, in general, more trusted than men (Shaub 1996; Wright and Sharp 1979).

[4]In our study, stimulus material was presented in the German language. Hence, the text length presented to our subjects was actually shorter than illustrated in Figure 2: CLAIM + QUALIFIER (8 w.) + DATA (14 w.) + BACKING (21 w.) + REBUTTAL (11 w.).

Pretest

Before conducting the fMRI study, we developed 104 different eBay offers (Figure 2 shows one example out of the entire set). We manipulated these offers with respect to the product description (based on Toulmin's model) and the picture of the product. Ten different USB flash drive images were randomly assigned to the various product descriptions. (The complete set of images is presented in the Appendix.) All other elements illustrated in Figure 2 were held constant throughout the study.

To guarantee variance in the trustworthiness of the eBay offers, 39 randomly selected subjects evaluated the developed offers on a scale ranging from 1 = "very low" to 7 = "very high." Afterward, we calculated the mean and standard deviation of each offer. To sharpen contrast, we selected the 10 most trustworthy offers (denoted as T, Trustworthy) and the 10 least trustworthy offers (denoted as U, Untrustworthy) for the fMRI experiment (means: T: 4.41 and U: 2.07).[5] An ANOVA confirmed the selected eBay offers as being significantly different concerning the level of trustworthiness (F (2,29) = 623.742, $p < .01$). *Post hoc* tests (Bonferroni, Tukey Multiple Comparison) showed significant results between the groups ($p < .01$).

[5]We also calculated the 10 neutral offers (N = 3.19). However, following an experimental procedure suggested by Stoll et al. (2008), the neutral offers were not included in the fMRI analysis.

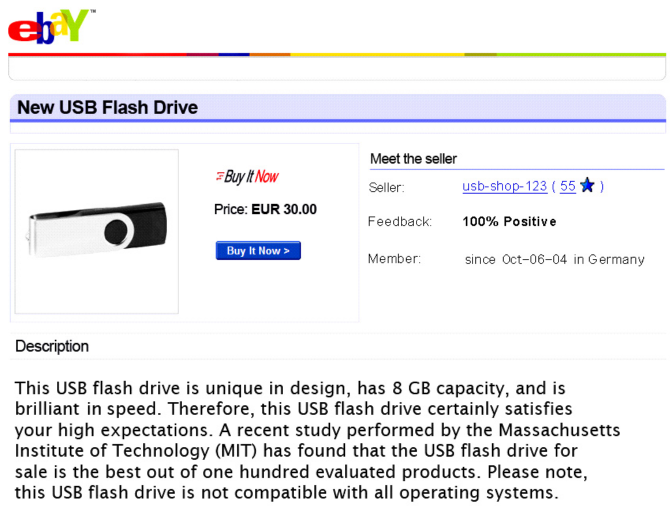


Figure 2. Example for a Trustworthy Internet Offer (eBay)

To ensure that during the fMRI experiment participants would not recognize that they were actually seeing the same eBay offer, we manipulated not only the product descriptions, but also the images of the USB flash drives. We used a nonparametric test (Kruskall-Wallis) to control for confounding influences within the groups T and U due to the images. The test showed no significant differences within each group (T: χ^2 (9) = 3.513, p = .940; U: χ^2 (9) = 3.748, p = .927). Consequently, the trustworthiness of the eBay offers was not affected by the images of the USB flash drives.

With respect to the influence of the product descriptions on the trustworthiness ratings, the pretest revealed that group U was solely comprised of eBay offers consisting of No Trust-Assuring Argument and Claim Only. In contrast, group T was comprised only of offers that consist of Claim + Data + Backing and Claim + Data + Backing + Rebuttal. Interestingly, our results show that Qualifiers had no influence on trustworthiness ratings.[6]

Main Study

Subjects

For the main study, we selected 10 female and 10 male subjects, all healthy and right-handed. We used fMRI as a method to identify possible gender differences in brain activation and subsequent trustworthiness ratings.[7] With respect to age, research indicates that trust increases almost linearly from early childhood to early adulthood, but stays relatively constant within different adult age groups (Phillips and Stanton 2004; Sutter and Kocher 2007). Hence, to avoid confounding effects due to age differences, we selected subjects from the narrow age group of 30 to 35 rather than using undergraduates or a blend of people from different age groups ($mean_{female}$ = 32.4, SD = 1.89, min = 30, max = 35; $mean_{male}$ = 31.2, SD = 1.39, min = 30, max = 34).

All participants were familiar with the Internet and eBay. The analysis revealed for duration of Internet usage: $mean_{female}$ = 81.6 months, SD = 41.49; $mean_{male}$ = 102.7 months, SD = 34.04; duration of eBay usage: $mean_{female}$ = 33.89 months, SD = 24.39; $mean_{male}$ = 55.56 months, SD = 36.86. An ANOVA showed no significant gender differences (duration of Internet usage: F (1,19) = 1.546, p = .23; duration eBay usage: F (1,17) = 2.162, p = .161).

We asked the participants whether the use of USB flash drives is important in their daily lives (five-point Likert scale with 1 = "totally disagree" and 5 = "totally agree"). As the results show, using USB flash drives is appreciably relevant for our subjects ($mean_{female}$ = 3.4, SD = 1.35; $mean_{male}$ = 3.5, SD = 1.27). An ANOVA showed no significant gender differences (F (1,19) = .029, p = .866).

[6]Class N included solely eBay offers that consist of Claim + Data. We did not use these offers in our fMRI analysis.

[7]Note that no subject of the fMRI experiment participated in the pretest.

To check for the general trust level regarding the brand eBay, or for an extremely positive or negative attitude toward the company, we asked the participants whether they generally have strong trust in eBay (five-point Likert scale with 1 = "totally disagree" and 5 = "totally agree"), and about their overall attitude toward eBay (five-point Likert scale with 1 = "extremely positive" and 5 = "extremely negative"). The analysis showed for trust in eBay: $mean_{female} = 2$, SD = .94; $mean_{male} = 3.2$, SD = 1.033; and for attitude toward eBay: $mean_{female} = 2.7$, SD = .67; $mean_{male} = 2.4$, SD = .69. An additional ANOVA showed specific gender differences for trust in eBay: $F(1,19) = 7.364$, $p = .014$, but not for the attitude toward eBay: $F(1,19) = .953$, $p = .342$.

Another important trait for which we assessed participants was the level of general trust (trust propensity). We measured general trust by a 25-item questionnaire (Rotter 1967).[8] At a maximum, each subject could score 125 points (high trust), and at a minimum, 25 points (low trust). Although the analysis showed no extreme outliers within the participants ($mean_{female} = 62.2$, SD = 7.97; $mean_{male} = 69$, SD = 7.16), the ANOVA revealed—as expected—a weak significant gender difference ($F(1,19) = 4.026$, $p = .06$). However, there were no arguments for excluding individual participants, because the calculated means for female and male participants are in line with the findings of similar studies (e.g., Kenning 2008).

Table 2 summarizes the descriptive statistics, revealing two important gender differences. First, women have significantly less trust in eBay than do men. Second, as expected, the level of general trust is significantly lower in women than in men. We refer to these differences in the "Discussion" section.

Regarding the fMRI experiment, standard exclusion criteria for MR examinations were applied (Savoy 2005). Because we employed visual stimuli, subjects with strong myopia or other relevant constraints of vision were excluded. All subjects provided written informed consent prior to the scanning sessions. The subjects were informed that the examination could potentially reveal medically significant findings, and they were asked whether they would like to be notified in such a case. An ethics commission approved the study. For their participation, all subjects received 20 Euro in cash and one of the ten USB flash drives (selected randomly).

Experimental Procedure and Stimulus Presentation

The eBay offers selected in the pretest were projected on a transparent screen with an LCD beamer, and viewed from the other side via a 45° mirror mounted on an element phase array coil. Furthermore, the offers were selected for equality in size, position, background, and luminance in order to prevent external confounding visual stimulation (Kenning et al. 2009). Each offer was visible for 12 seconds. The task for the participants was to press one of two corresponding buttons on a magnetic resonance compatible response box to indicate, at the end of the 12-second time frame, whether they considered an offer to be trustworthy or untrustworthy. After the 12-second offer image, participants saw a fixation cross for 3 seconds. Then the next offer was presented, and the displays continued in this way. The sequence of the offers was pseudo-randomized for every subject. In total, every subject had to evaluate 120 offers.

Data Collection and Analysis

The study was executed on a 3T fMRI-scanner (Magnetom Trio, SIEMENS, Erlangen, Germany).[9] fMRI is a noninvasive neuroscientific technique with relatively good spatial and temporal resolution (Huettel et al. 2009). It takes advantage of the blood oxygenation level dependent (BOLD) effect for estimating the neural activity that corresponds with the experimental conditions. The data set used in this study consisted of 36 transversal slices of 3.6 mm thickness without a gap, a field of view of 230 mm × 230 mm, an acquired matrix with 64 × 64 (i.e., isotropic voxels with 3.6 mm edge length). Contrast parameters were a signal response time of 3000 ms, echo time of 50 ms, and a flip angle of 90°.

Data analysis was conducted with the SPM5-freeware (Friston 1996; Friston et al. 1995) using MatLab as a working base, the application followed procedures described in Kenning et al. (2007) and Poldrack et al. (2007). The data preprocessing consisted of three initial steps. First, to correct for artifacts due to participant head movement in the scanner, all images were realigned by a "rigid body" transformation to the first image of the session (realignment). Second, to compare all participants within the group analysis, all images were normalized and re-sampled to the standard Montreal Neurological Institute (MNI) template (normalization). Third, to prepare the data for the statistical analysis, all images were smoothed with a 5-mm Gaussian kernel (Ashburner et al. 1997).

[8]For a recent review on findings yielded by the Rotter scale, see Cook et al. (2005).

[9]The protocol included a 3D isotropic T1-weighted data set of the whole head, with a measured voxelsize of 1.0 mm edge length for anatomical identification and coregistration into the Talairach-Space. Functional images were acquired using a T2* weighted single-shot gradient echo-planar imaging (EPI) sequence, which covered nearly the entire brain.

Table 2. Descriptive Statistics			
	Women (*N* = 10) Mean (SD)	Men (*N* = 10) Mean (SD)	ANOVA (Gender Differences)
Age (in Years)	32.4 (1.89)	31.2 (1.39)	F (1,19) = 2.592 p = .125
Duration of Internet Usage (in Months)	81.6 (41.49)	102.7 (34.04)	F (1,19) = 1.546 p = .23
Duration of eBay Usage (in Months)	33.89 (24.39)	55.56 (36.86)	F (1,17) = 2.162 p = .161
Importance of USB Flash Drives in Daily Life[a]	3.4 (1.35)	3.5 (1.27)	F (1,19) = .029 p = .866
Trust in eBay[a]	2 (.943)	3.2 (1.03)	F (1,19) = 7.364 p = .014*
Attitude toward eBay[b]	2.7 (.67)	2.4 (.69)	F (1,19) = .953 p = .342
General Trust[c]	62.2 (7.97)	69 (7.16)	F (1,19) = 4.026 p = .06*

Notes: *Significant gender difference.
[a]Five-point Likert scale with 1 = "totally disagree" and 5 = "totally agree."
[b]Five-point Likert scale with 1 = "extremely positive" and 5 = "extremely negative."
[c]Rotter (1967) scale with 25 = minimum trust and 125 = maximum trust.

For the estimation of the general linear model (GLM), onsets were constructed that constitute a matrix that included one vector for each trustworthiness level (T, U). Realignment parameters were included as additional covariates. The estimation of the GLM was conducted by fitting a reference hemodynamic response function to each event (onset) in the observed data. In order to investigate gender- and trust-dependent activity changes, one- and two-sample t-tests were accomplished on the second-level (group) analysis. All coordinates received by SPM5 were transformed to the Talairach and Tournoux space and were assigned to cortical regions with the T2T-database Java applet.[10]

Results

Behavioral Results

Before we were able to use the two trustworthiness groups from the pretest (T, U), we had to control the evaluation of trustworthiness of the fMRI participants. The analysis showed congruent results between the pretest and the fMRI study. Therefore, as summarized in Table 3, the group classification was confirmed for both female (F (1,79) = 160.17, p < .001) and male participants (F (1,79) = 217.754, p < .001).

Our results show significant gender differences regarding the trustworthiness evaluation of eBay offers. That is, in the T group, on average, women rated 68.5 percent of the eBay offers as trustworthy, whereas only 57.7 percent of male participants, on average, did so (F (1,79) = 6.985, p < .01). In the U group, on average, women rated 20.0 percent of the eBay offers as trustworthy, whereas only 11.0 percent of male participants, on average, did so (F (1,79) = 9.903, p < .001). Consequently, results show that women evaluated the eBay offers significantly higher in trustworthiness than did men. This result is surprising, because the female participants exhibited a lower level of trust in eBay, as well as a lower level of general trust (see Table 2). We refer to this result in the "Discussion" section.

Neuroimaging Results

By contrasting the brain responses of the female and male participants, statistical parametric maps were generated that displayed the t-value of each voxel that reached a p < .001

[10]The applet is available at http://wwwneuro03.uni-muenster.de/ger/t2tconv/.

Table 3. Evaluation of Trustworthiness During the fMRI Experiment

Group		Women (N = 10)	Men (N = 10)	ANOVA (Gender Differences)
Trustworthy eBay offers	Mean	68.5	57.7	F (1,79) = 6.985 $p < .01$
	SD	19.2	17.0	
	Median	70.0	60.0	
Untrustworthy eBay offers	Mean	20.0	11.0	F (1,79) = 9.903 $p < .001$
	SD	14.6	10.5	
	Median	20.0	10.0	
	ANOVA (Group Differences)	F (1,79) = 160.17 $p < .001$	F (1,79) = 217.754 $p < .001$	

[uncorrected] with an extent threshold of five voxels.[11] We found significant variance in brain activation patterns between women and men, yet also a few similarities (see Table 4).

Trustworthy Versus Untrustworthy Offers

We found significantly higher brain activation in women in the thalamus, striatum (putamen), and fusiform gyrus (BA 37). In contrast, we found higher activation in men only in the DLPFC (BA 9). Interestingly, we noted increased activation in the dorsal ACC (BA 32) in both women and men, although the cluster size (in voxels) was much larger in men (see Table 4 and Figure 3). Moreover, we found increased activation in the lingual gyrus and cuneus (BA 18) for both genders. However, because these two brain regions are not trust-specific—we assume that they are a result of the nature of the experimental task, that is, reading product descriptions (Hagoort et al. 1999; Mechelli et al. 2000; Price et al. 1996)—we do not discuss them in this article.

[11]To identify brain activity changes during certain tasks, it is necessary to construct contrasts in the SPM (statistical parametric mapping) tool. The contrast trustworthy versus untrustworthy offers is defined as +1 (trustworthy) and -1 (untrustworthy) and uses the subtraction method from set theory. Thus, SPM subtracts the neural activity during the processing of untrustworthy offers from the activity during the processing of the trustworthy offers. What remains is the neural activity from the trustworthy offers, which is significantly different from the neural activity of the untrustworthy offers. The contrast untrustworthy versus trustworthy offers is defined as +1 (untrustworthy) and -1 (trustworthy). The neural activity that results from the processing of the trustworthy offers is subtracted from the activity resulting from the processing of the untrustworthy offers. What remains is the positive difference between untrustworthy and trustworthy offers.

Untrustworthy Versus Trustworthy Offers

We found significantly higher brain activation for women in the ventral ACC (BA 24), hippocampus, DLPFC (BA 9), and caudate nucleus (body). In contrast, we found higher activation for men in the VMPFC (BA 10) and ventral posterior cingulate cortex (BA 23). Moreover, we found increased activation in the insular cortex, with exactly the same cluster size, in both women and men (see Figure 4 for the female activation).

In *Hypothesis 1*, we predicted that decision making on both trustworthy and untrustworthy Internet offers activates *different brain areas* in men and women. Our results support this hypothesis (see Table 4).

With respect to our *Hypotheses 1.1* through *1.7*, which predicted gender differences in the activation in seven trust-specific brain regions, we found full support for five of them—ACC (1.1), thalamus (1.2), hippocampus (1.3), VMPFC (1.5), and striatum (1.6). For one hypothesis—DLPFC (1.4)—we found partial support. For another hypothesis—insular cortex (1.7)—we did not find support.

In *Hypothesis 1.4*, we predicted that decision making on both trustworthy and untrustworthy Internet offers activates the DLPFC significantly more in men than in women. We found activation in the DLPFC in men, but also in women (Table 4). In *Hypothesis 1.7*, we predicted that decision making on untrustworthy Internet offers activates the insular cortex significantly more in women than in men. We found activation in both the female and male insular cortex. We refer to these results in the "Discussion" section.

In *Hypothesis 2*, we predicted that decision making on both trustworthy and untrustworthy Internet offers activates *more brain areas* in women than in men. Our results support this

Table 4. Brain Areas Activated During the Presentation of Internet Offers

		Cluster Size (Voxels)	Laterality (Left/Right)	MNI-Coordinates[a] (x, y, z)		
Trustworthy Versus Untrustworthy Offers ([T] – [U])						
Women (*N* = 10)						
Gender Difference	Thalamus	6	R	10	-16	8
	Striatum (Putamen)	5	R	22	20	-2
	Fusiform Gyrus (BA 37)	8	L	-38	-58	-16
Gender Similarity	Dorsal Anterior Cingulate Cortex (BA 32)	5	L	-4	8	26
	Lingual Gyrus[b]	34	R	52	28	16
	Cuneus (BA 18)[b]	30	R	2	-70	42
Men (*N* = 10)						
Gender Difference	Dorsolateral Prefrontal Cortex (BA 9)	25	R	38	32	26
Gender Similarity	Dorsal Anterior Cingulate Cortex (BA 32)	138	R	8	14	57
	Lingual Gyrus[b]	42	L	-8	-90	16
	Cuneus (BA 18)[b]	7	L	-10	-72	12
Untrustworthy Versus Trustworthy Offers ([U] – [T])						
Women (*N* = 10)						
Gender Difference	Ventral Anterior Cingulate Cortex (BA 24)	15	L	-2	38	2
	Hippocampus	12	R	30	-44	-2
	Dorsolateral Prefrontal Cortex (BA 9)	24	L	-18	46	34
	Caudate Nucleaus (Body)	31	L	-12	12	16
Gender Similarity	Insular Cortex	10	L	-40	-16	-4
Men (*N* = 10)						
Gender Difference	Ventromedial Prefrontal Cortex (BA 10)	18	L	-10	46	8
	Ventral Posterior Cingulate Cortex (BA 23)	45	L	-4	-56	18
Gender Similarity	Insular Cortex	10	L	-46	-8	0

Notes: T: Trustworthy eBay offer. U: Untrustworthy eBay offer. BA: Brodmann Area. MNI: Montreal Neurological Institute.
[a]The coordinates of the peak of the activation are given according to MNI-space (Montreal Neurological Institute).
[b]Brain areas are not trust-specific. Rather, we assume that they are a result of the nature of the experimental task (i.e., reading product descriptions) (Hagoort et al. 1999; Mechelli et al. 2000; Price et al. 1996).

hypothesis. Table 4 shows that the contrast of trustworthy versus untrustworthy offers resulted in significant activity changes in *four female* and *two male* brain areas (without considering the lingual gyrus and cuneus, activation of which is not believed to be trust-specific). When contrasting untrustworthy versus trustworthy offers, we found activity changes in *five female* and *three male* brain areas.[12]

Before we discuss our findings regarding the brain activation patterns, it is useful to briefly link our results to previous neuroscience studies in the trust realm. These studies were either conducted in the context of economic games (Baumgartner et al. 2008; Delgado et al. 2005; King-Casas et al. 2005; Krueger et al. 2007) or they concerned the assessment of the trustworthiness of human faces (Winston et al. 2002).[13]

[12]In Table 4, we also report the cluster size (in voxels) and laterality of the various brain activations. However, an in-depth discussion of gender-specific cluster size and laterality issues is not feasible within the scope of the present article. Recent brain research gives insights into this topic (e.g., Cahill 2006; Cosgrove et al. 2007). The topic is a possible avenue for future research.

[13]In most studies, the economic game investigated is the trust game, which has been developed to operationalize trust and trustworthiness (Berg et al. 1995). In a trust game, two players, X (the investor) and Y (the trustee), each have an initial endowment. First, X decides whether to keep his endowment or to send (a part of) it to Y. Then Y observes X's action and decides whether to keep the amount she received or share some of it with X. The experimenter triples X's transfer, so that both players are better off collectively if X transfers money and Y sends back a sufficient amount. This situation mimics a sequential economic exchange in the absence of contract enforcement institutions. Y has a strong incentive to keep all the money and repay none to X. If X anticipates this behavior, however, there is little reason to transfer. Consequently, if X transfers no money, then a chance for mutual gain would be lost. In the trust game, the amount sent by X is used as a quantitative measure for trust, and Y's transfer back is used as a measure for trustworthiness. It should be noted that a recent paper challenges the assumption that the amount sent by X is a reliable measure for trust (Kugler et al. 2009).

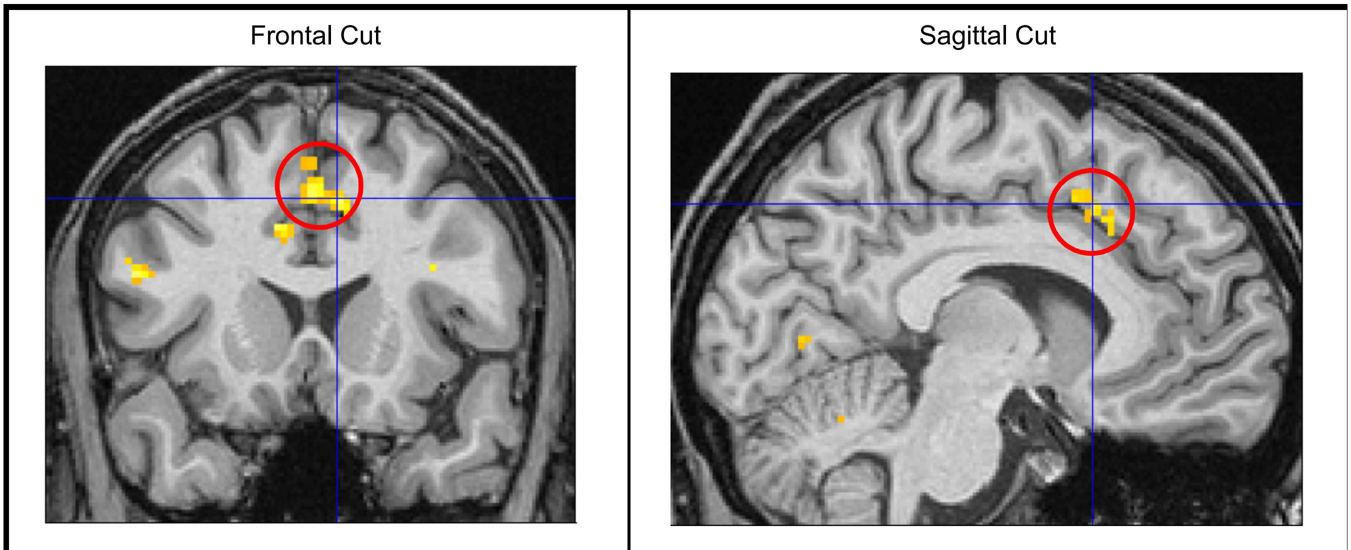


Figure 3. Male Dorsal Anterior Cingulate Cortex (BA 32), Trustworthy Versus Untrustworthy Offers

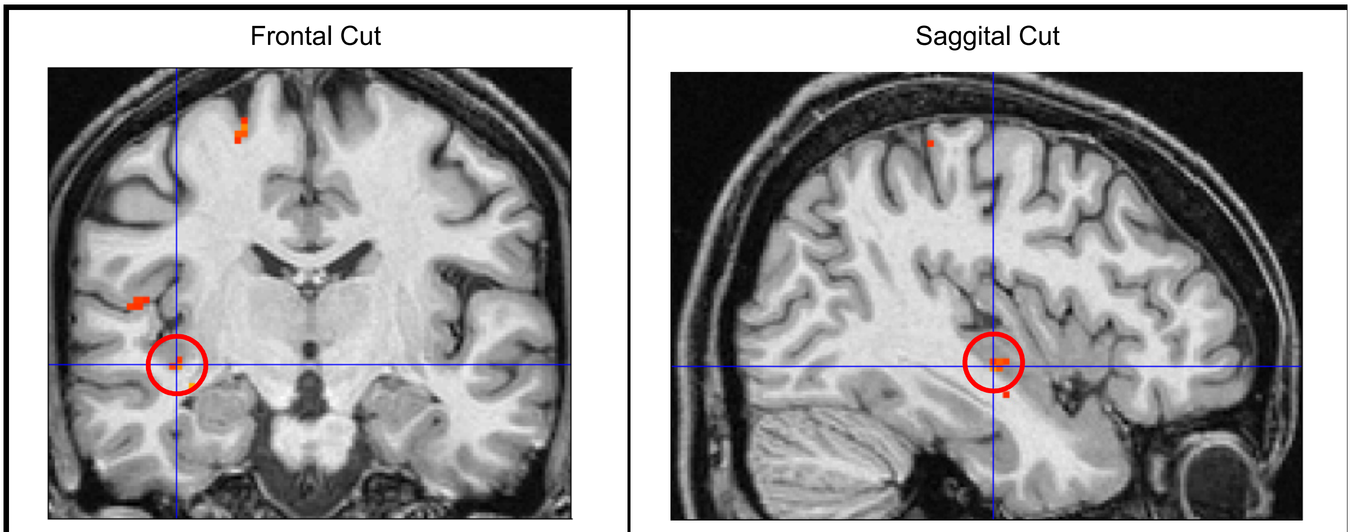


Figure 4. Female Insular Cortex, Untrustworthy Versus Trustworthy Offers

Table 5. fMRI Studies on Human Trusting Behavior and Associated Brain Regions

	Baumgartner et al. (2008) (Trust Game)	Delgado et al. (2005) (Trust Game)	King-Casas et al. (2005) (Trust Game)	Krueger et al. (2007) (Trust Game)	Winston et al. (2002) (Trustworthiness of Human Faces)	Present Study (Trustworthiness of Internet Offers)
Sample Size (Female/Male)	**49 (0/49)**	**12 (5/7)**	**96 (n.a.)**	**44 (22/22)**	**14 (6/8)**	**20 (10/10)**
Amygdala	•				•	
Ventral Anterior Cingulate Cortex (BA 24)	•		•			F
Caudate Nucleus	•	•	•			F
Dorsolateral Anterior Cingulate Cortex (BA 32)		•		•		F/M
Dorsolateral Prefrontal Cortex (BA 9)				•		F/M
Hippocampus and Parahippocampus Gyrus		•			•	F
Insular Cortex	•	•			•	F/M
Putamen	•	•				F
Thalamus	•		•		•	F
Ventral Posterior Cingulate Cortex (BA 23)						M
Ventromedial Prefrontal Cortex (BA 10)						M

Notes: The table lists brain regions that are mentioned in at least two of the six studies. Exceptions are the ventral posterior cingulate cortex (BA 23) and the ventromedial prefrontal cortex (BA 10), which were identified in the present study. Only the present study reports gender-specific results. BA: Brodmann Area. F: Female. M: Male. n.a.: not available.

Table 5 summarizes these investigations, thereby supporting evaluation of whether or not the previous findings are in line with the results of the present study. A marked cell in Table 5 indicates that a study has found this brain region activated in trust-related tasks. Except for the activity results in the amygdala, our study confirms previous research findings.[14] However, although most of the previous studies have used a mixed-gender sample (with the exception of the study of by Baumgartner et al., and with information unavailable in King-Casas et al.), no available study reports specifically on gender differences in brain activation. Therefore, our study makes a substantial contribution to a greater understanding of neuronal gender differences in trust research.

Discussion

To show empirically that decision making on both trustworthy and untrustworthy Internet offers is associated with gender-specific activity changes in certain brain areas, we used fMRI.Altogether, when we contrasted trustworthy versus untrustworthy offers, and untrustworthy versus trustworthy

[14]The amygdala was only marginally activated when the participants were exposed to untrustworthy Internet offers (< 5 voxels, p < .001). Lesion studies conducted by the research group of Adolphs, Damasio, and Tranel might explain this finding (Adolphs, Tranel, and Damasio 1998; Adolphs et al. 2005). They conducted three case studies and found that patients with complete bilateral amygdala damage judged other people to look more trustworthy and more approachable than did normal viewers or patients with brain damage in other areas, providing empirical evidence that the amygdala is associated with distrust. Moreover, in their study they also asked the three patients to rate the likeability of different individuals based on single words (i.e., adjectives describing the people) presented visually on a sheet of paper. Intriguingly, the study found that the subjects' impairment does not extend to judging verbal descriptions of people. As a result, Adolphs and his colleagues conclude that the amygdala is necessary to judge the trustworthiness of faces, but it is not necessary to evaluate the trustworthiness of sentences and words. The results of our fMRI study substantiate this theoretical perspective, which has also been confirmed by recent work of Todorov and his colleagues (e.g., Engell et al. 2007; Todorov 2008; Todorov et al. 2008). Additionally, Bechara and Damasio (2005, p. 353) state that the amygdala has evolved for a survival purpose, for example, to be responsive to dangerous animals or persons with untrustworthy faces. Therefore, it might be possible that the role of the amygdala in today's economic decision situations is not the same as it was in former times, because Internet transactions do not require both sellers and buyers to reveal their faces. Rather, conclusions about a transaction partner's characteristics have to be drawn on the basis of other factors, such as feedback profiles (Ghose et al. 2005; Pavlou and Dimoka 2006) or product description texts, as discussed in the present article.

offers, we found some similarities and substantial differences between neural processing in women and men (Table 4). Our study thereby reveals considerable gender differences in neurobiology during decision making on Internet offers, which supports our theorizing in *Hypothesis 1*.

As evidence in support of *Hypothesis 2*, we recognize the significant result that women, in contrast to men, activated more brain areas during the presentation of both trustworthy and untrustworthy Internet offers (Table 4). This result replicates findings from other neuroscience studies in various domains. One study (Canli et al. 2002), for example, investigated gender differences in brain activation patterns of emotional memories and found that women recruited more brain regions than men. In another study (Kucian et al. 2005), it was observed that women activated more brain areas than men in a calculation and mental rotation task. The present study, therefore, extends to the online trust domain the existing finding that women, in contrast to men, activate more brain areas in cognitive tasks (e.g., memory, mathematics, or decision making).

In purchasing online, women, in general, perceive a higher level of risk than do men (e.g., Garbarino and Strahilevitz 2004). Moreover, the descriptive analysis of our survey data shows that women have both less general trust and less trust in eBay (Table 2). To overcome this increased risk perception and the reduced levels of general trust and trust in eBay, elevated levels of information search, and of consideration of that information, could be useful, because this can lead to a higher level of confidence in a decision (Hall et al. 2007). However, given that such deliberate search for and consideration of increased information should be reflected in a more sophisticated brain processing pattern, women's inclination toward a heightened level of risk perception may explain why more brain areas are activated in women than in men.

In the following section, we provide additional discussion of our findings, which outlines gender differences and similarities in brain activation.

Gender Differences

Trustworthy Versus Untrustworthy Offers

By contrasting trustworthy versus untrustworthy offers (upper part in Table 4), we found increased activation

- in the female thalamus, striatum (putamen) and fusiform gyrus (BA 37), and
- in the male DLPFC (BA 9).

With respect to the gender differences in *thalamus*, *DLPFC*, and *striatum (putamen)* activation, these results support our theorizing in *Hypotheses 1.2*, *1.4*, and *1.6*. Moreover, we confirm findings from other neuroscience studies which have also found that these brain regions are activated in trust situations (Baumgartner et al. 2008; Krueger et al. 2007).

In the trust literature, the *fusiform gyrus* is mentioned in only one study (Delgado et al. 2005), and even in this study the area is not discussed as a major trust region. Therefore, we have not formulated a hypothesis about this brain region. However, research has determined an important role of the fusiform gyrus in word and picture processing (e.g., Starrfelt and Gerlach 2007). Therefore, it is possible that this brain region was activated as a result of the experimental task (i.e., visually processing Internet offers). Although recent research (Salloum et al. 2008) has studied gender differences in fusiform gyrus activation in a cognitive task, further research is needed to investigate its role in trust situations and corresponding gender differences.

Untrustworthy Versus Trustworthy Offers

By contrasting untrustworthy versus trustworthy offers (lower part in Table 4), we found increased activation

- in the female ventral ACC (BA 24), hippocampus, DLPFC (BA 9), and caudate nucleus (body), and
- in the male VMPFC (BA 10) and ventral posterior cingulate cortex (BA 23).

With respect to the gender differences in *ventral ACC*, *hippocampus*, *VMPFC*, and *striatum (caudate nucleus, body)* activation, the results support our theorizing in *Hypotheses 1.1*, *1.3*, *1.5*, and *1.6*. Moreover, activation in these four brain areas confirms the findings from other neuroscience studies (Baumgartner et al. 2008; Delgado et al. 2005; King-Casas et al. 2005; Winston et al. 2002). The activation in the male VMPFC is illustrated in Figure 5.

Interestingly, we found increased activity in the *female striatum* (i.e., caudate nucleus, body) during the presentation of untrustworthy offers, as shown in Figure 6. Caudate activation in women conforms to the well-known anatomical fact that the female caudate is, on average, larger than that of the male (Cosgrove et al. 2007). In this case, however, no previous studies showing activation in the caudate nucleus during the presentation of untrustworthy stimuli could be found. We believe that the activation in the female striatum during the presentation of both trustworthy and untrustworthy offers is partly associated with the highly positive attitudes

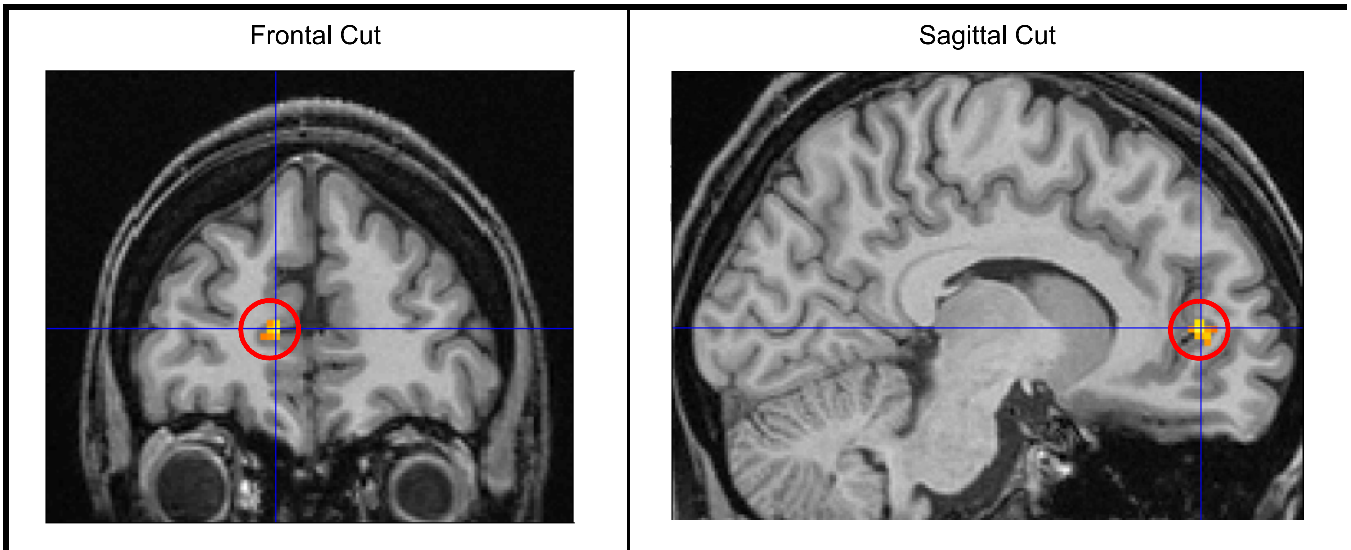


Figure 5. Male Ventromedial Prefrontal Cortex (BA 10), Untrustworthy Versus Trustworthy

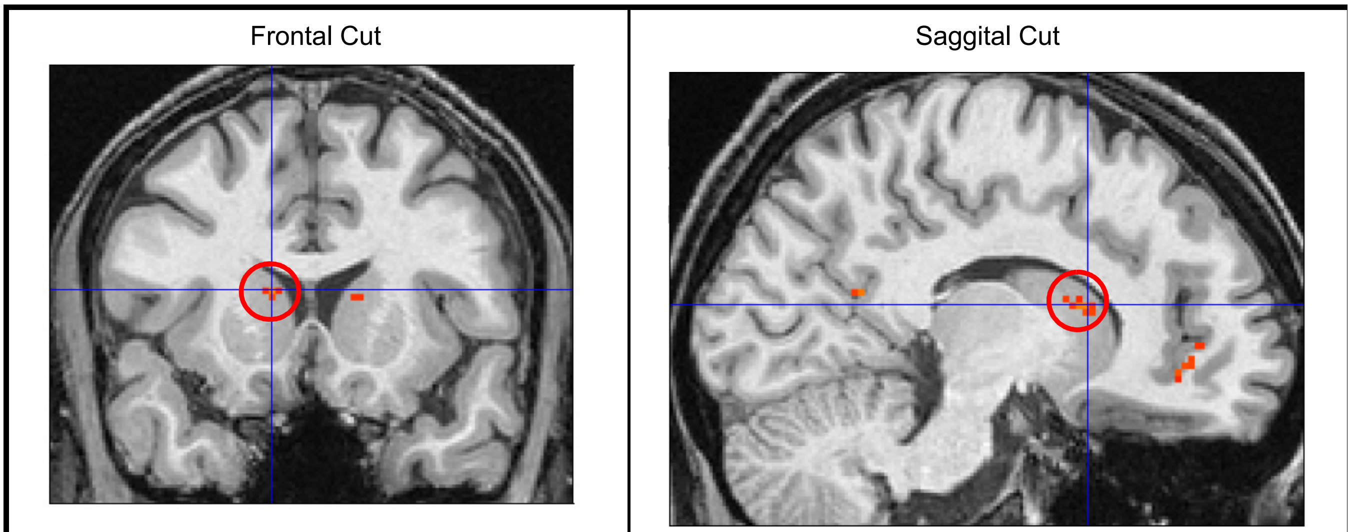


Figure 6. Female Caudate Nucleus (Body), Untrustworthy Versus Trustworthy Offers

toward buying held by women generally (e.g., Campbell 2000). In essence, women tend to focus on the often enjoyable *process* of buying ("shopping feeling") rather than simply on the *outcome*.

In our presentation of gender differences in online trust and related behaviors, we noted that the findings of a study by Dittmar et al. (2004) challenge those of Swaminathan et al. (1999). Specifically, we noted that additional empirical evidence is needed to clarify whether socio-psychological and emotional concerns remain most important for women in online shopping, or whether functional factors become the more relevant category. Our fMRI data (i.e., the striatum activation) provide empirical support that women consider socio-psychological and emotional concerns most relevant in both conventional shopping and online shopping.

Moreover, our behavioral results show that although women have less trust in eBay and a lower level of general trust than do men (see Table 2), their trust in both trustworthy *and* untrustworthy offers is significantly higher than men's trust (see Table 3). The argument that women generally have a more positive attitude toward buying than men (shopping feeling), and as a result show higher activation in the striatum, is a possible explanation for this paradoxical finding. Overall, our data suggest that in shopping situations, women, in general, may have a stronger tendency to activate the striatum (a reward processing area) than men.

Interestingly, a recent study on gender differences in brain activation during computer game playing (Hoeft et al. 2008) found that men, but not women, strongly recruit the striatum. Therefore, men rather than women perceive game playing as rewarding. Considering this finding together with our results, we believe that there is converging evidence that demonstrates that the same task (e.g., either shopping or computer game playing) is perceived differently by women and men, in particular, due to activation differences in the brain's reward area, the striatum.

In *Hypothesis 1.4*, we predicted that decision making on both trustworthy and untrustworthy Internet offers activates the DLPFC significantly more in men than in women. We found support for this prediction during the presentation of trustworthy Internet offers (upper part in Table 4), but we did not find support for this prediction during the presentation of untrustworthy Internet offers (lower part in Table 4). DLPFC was shown to play a central role in coupling the information of rewards to actions, thereby controlling behavior and optimizing decision making (Heekeren et al. 2006; Lee and Seo 2007; Van't Wout et al. 2005). Because coupling information about rewards to actions (and thereby controlling behavior and optimizing decision making), is a cognitive rather than affective process, which empathizing–systemizing theory suggests will be observed in men rather than women (Baron-Cohen et al. 2005), we hypothesized that the DLPFC should be activated significantly more in men than in women.

However, in a study of the neuronal activity of players in an economic game, Sanfey et al. (2003) found increased activity in the DLPFC as a response to unfair offers. Therefore, it is possible that women, but not men, considered untrustworthy offers as unfair offers, which could be associated with the perception of uncertainty and risk. This theorizing is in line with women's increased inclination toward risk perception in online shopping (e.g., Garbarino and Strahilevitz 2004).

In the trust literature, the *ventral posterior cingulate cortex (BA 23)* has not yet been identified as a relevant brain region. We have not, therefore, formulated a hypothesis about this brain area. However, although the posterior cingulate cortex (PCC) has not been studied as intensively as the ACC, its critical role in perception and decision making is established. One study (Maddock et al. 2003), for instance, found that activation in the PCC is associated with emotionally arousing stimuli. Moreover, an animal study (McCoy and Platt 2005) found that neuronal activity in PCC increased when monkeys made risky choices, and scaled with the degree of risk. Due to the interrelationship between risk and trust (e.g., Pavlou 2003), our data suggest that in untrustworthy situations PCC activity is likely to signal both arousal and risk in men, but not in women. Recent research has found that hormonal differences between women and men influence receptor systems in the PCC (Cosgrove et al. 2007), a result that is in line with the gender differences found in the present study.

Gender Similarities

Trustworthy Versus Untrustworthy Offers

By contrasting trustworthy versus untrustworthy offers, we found increased activity in the *dorsal ACC* in both women and men, with a larger cluster size in men (Table 4). In *Hypothesis 1.1*, we predicted that decision making on both trustworthy and untrustworthy Internet offers activates the dorsal ACC (BA 32) in men and the ventral ACC (BA 24) in women. This prediction was based on evidence which shows that, at a population level, men are stronger systemizers (implying more cognitive information processing, which was found to occur in the dorsal ACC) and women are stronger empathizers (implying more affective information processing, which was found to occur in the ventral ACC) (Baron-Cohen et al. 2005; Bush et al. 2000; Devinsky et al. 1995).

In our study, we found that the dorsal ACC was activated not only in men (as predicted), but also in women (Table 4). At a first glance, this result seems to contradict our theorizing in *Hypothesis 1.1*. However, because the cluster size of dorsal ACC activation in men is much larger than in women (Table 4), we believe that this gender difference supports the notion that men, in contrast to women, usually process information in a cognitive rather than affective manner (Cosgrove et al. 2007; Meyers-Levy 1994; Wager et al. 2003). Therefore, our result does not contradict the empathizing–systemizing theory.

Untrustworthy Versus Trustworthy Offers

By contrasting untrustworthy versus trustworthy offers, we found increased activity in the *insular cortex* in both women and men, with an equal cluster size (Table 4). This finding does not support the hypothesized gender difference in insular activation (*Hypothesis 1.7*). Women, in general, perceive greater risks in a wide variety of domains, including environmental, financial, and medical (e.g., Byrnes et al. 1999), as well as Internet shopping (Garbarino and Strahilevitz 2004; Sheehan 1999). Consequently, we predicted that the insular cortex, a brain area responsible for encoding uncertainty and risk (e.g., Krain et al. 2006), is activated significantly more in women than in men during the presentation of untrustworthy Internet offers. However, because there is no gender difference in insular cortex activation, other brain areas listed in Table 4 must account for the behavioral gender differences in uncertainty and risk perception. In the present study, in particular, we identified the hippocampus as a brain region that signals uncertainty and risk in women, but not in men (Table 4).

The present study identifies the insular cortex as a brain area that is important for both women and men reacting to untrustworthy Internet offers. In the following, we discuss insular activation in the context of *situational normality*, an important behavioral construct in IS trust research (e.g., Gefen et al. 2003), thereby outlining an example of neuroscience findings on human behavior that could be effectively integrated into IS research.

Buying and selling online is inherently different from traditional commerce. First, buyer and seller are physically separated, permitting simultaneous exchange of product and money. Second, a typical online transaction takes place between a buyer and a seller who are unknown to each other, on the basis of a one-time interaction (Resnick and Zeckhauser 2002). Third, in Internet transactions, the medium of communication consists of a complex system of technical components that are not well understood by most users. Fourth, the Internet transaction is influenced by the fact that facial expression, gesture, and similar cues are not available (Brosig et al. 2003). From a buyer's perspective, these characteristics of online transactions may lead to considerable perceptions of uncertainty and risk, both among female and male users (Gefen et al. 2008; Kim et al. 2009; McKnight et al. 2002a; Pavlou 2003; Pavlou et al. 2007).Hence, our results concerning the activation of the insular cortex are in line with behavioral research.

In particular, we believe that insular activation may be associated with situational normality, which signifies a condition in which a person believes that the situation in a risky venture is in proper order and is favorable, and as a result expects that success is likely and failure is unlikely (Baier 1986; Lewis and Weigert 1985).[15] An online buyer who perceives high situational normality would believe that the Internet environment is appropriate, fitting, and favorable for engaging in a transaction. Situational normality, therefore, assures people that everything in the online environment is as it ought to be, and that a shared understanding of what is happening exists (McKnight and Chervany 2001; McKnight et al. 1998; Zucker 1986).

People tend to perceive greater trust when the nature of the interaction is in accordance with what they consider to be typical and, hence, anticipated. Gefen et al. (2003) point out that this is in accord with sociologists such as Luhmann (1979) and Blau (1964), who view trust as the product of fulfilled expectations. In the context of Internet offers, this view is relevant in that an offer represents what customers expect, based on their experience and knowledge of other similar offers. If the expectations are met, then a buyer will be more inclined to trust the seller. However, research suggests that a state of distrust is the mental system's signal that the environment is not normal (Schul et al. 2004, 2008).[16] To be able to recognize that an environment is not normal, experience with that environment is necessary. When exposed to a known environment, it is very likely that humans retrieve their previous experience with that environment. If the perceived environment does not meet the expectations, then distrust is likely to emerge; if it meets the expectations, then distrust is unlikely and an essential basis for the emergence of trust does exist.

Accordingly, the activation of the insular during the presentation of untrustworthy offers can be interpreted as a neuronal signal for distrust. The participants in our fMRI study are highly experienced Internet and eBay users (see Table 2). As a result, it is likely that the participants retrieved their experience with eBay during the visual perception of the stimuli in order to detect abnormalities. Because technical products (e.g., digital cameras, laptops, or USB flash drives) are often described at length on eBay (Bolton et al. 2004; Gregg and Scott 2008), the "abnormality hypothesis" might

[15]Situational normality is discussed as an important antecedent of institution-based trust in the IS literature (e.g., Gefen et al. 2003; McKnight and Chervany 2001; McKnight et al. 1998).

[16]It must be noted that the focus of the present study is not to explicitly distinguish between trust and distrust. However, there is increasing evidence that both constructs are not the two ends of one continuum; rather, they are two separate constructs (e.g., Cho 2006; Lewicki et al. 1998; McKnight and Chervany 2001; McKnight and Choudhury 2006; McKnight et al. 2002b; McKnight et al. 2004; Ou and Sia 2009). Brain imaging could be used to show that trust and distrust are in fact two separate constructs on the neurological level of analysis.

explain why the conditions NO TRUST-ASSURING ARGUMENT or CLAIM ONLY—which included either little or no text—were rated as untrustworthy. Detecting and avoiding threats are two of the most basic human skills, so that the sensitivity to risks of the insular is, in both men and women, considered to be a mechanism that promotes learning in order to avoid loss (Samanez-Larkin et al. 2008).[17]

Summary, Implications, Limitations, and a Concluding Comment

One major goal of our study was to add a new theoretical perspective to the existing IS trust literature. Taking into account the mounting evidence on gender differences in online trusting behavior, we set up a laboratory experiment to explore whether or not these differences are accompanied by neural differences that can be detected by fMRI. In fact, we found a considerable number of neural differences. The following discussion summarizes our results and outlines important implications.

1. We found that there are several gender-specific *trust* regions (Table 4), but there are also two gender-independent trust areas (dorsal ACC and insular cortex). Trustworthy offers are associated with female brain activation in the dorsal ACC, thalamus, striatum (the putamen), and fusiform gyrus (BA 37), whereas in the male brain the dorsal ACC and DLPFC (BA 9) are activated. Untrustworthy offers, in contrast, activated the female ventral ACC, hippocampus, DLPFC, striatum (the caudate body), and insular cortex; in men, the untrustworthy offers activated the VMPFC (BA 10), insular cortex, and ventral PCC (BA 23). Altogether, our study shows that women recruit more limbic structures than do men in a trustworthiness evaluation task.

2. In the "Literature Review and Hypotheses" section and in the "Discussion" section, we link our results to the literature in both IS and neuroscience, thereby providing information about the gender-specific *localization* and *nature* of trustworthiness. For example, we discuss the potential role of the insular in abnormality detection, linking our discussion to both the IS trust literature (e.g., situational normality is discussed in McKnight and Chervany 2001; McKnight et al. 1998; Gefen et al. 2003) and the (social) neuroscience trust literature (Baumgartner et al. 2008; Delgado et al. 2005; King-Casas et al. 2005; Krueger et al. 2007). In our research, we found that the presentation of both trustworthy and untrustworthy Internet offers triggered activation in more brain areas for females than for males, thereby providing neurological evidence for the commonly accepted view that, in general, women process more information, and do so more comprehensively, than men (e.g., Meyers-Levy and Maheswaran 1991). This pattern of female brain activation may occur because women perceive greater levels of uncertainty and risk (e.g., Byrnes et al. 1999).

3. We demonstrated that product descriptions (such as those we embedded in an eBay site) can trigger different brain activation in both women and men, thereby identifying an important *antecedent* of the trustworthiness of Internet offers (Table 4). Our study shows that similar descriptions are processed differently by women and men, which is a finding with considerable implications for online shop managers, engineers, and marketers. In particular, trust-building strategies should differentiate between women and men. This differentiation could, for example, address presentation format, content, and color (Simon 2001; Stenstrom et al. 2008). A related and recent study by Cyr et al. (2009) investigated whether *human images* with facial features would induce a user to perceive a Web site as more trustworthy, but did not find a direct relationship. However, our study revealed that *text* has the potential to induce trustworthiness. Replicating the study of Cyr et al. by means of fMRI would be a useful avenue for future research, because brain activation analysis has the potential to disclose possible reasons for a lack of direct relationship between human images and trustworthiness. For example, the fusiform gyrus is a brain area well known for the processing of human faces (Kanwisher and Yovel 2006; Starrfelt and Gerlach 2007), and holds prospects for developing brain activity studies using fMRI.

4. We showed that increased activation in specific brain regions correlates with women's and men's trustworthiness ratings, thereby providing evidence for the *relationship* between brain activation and behavior (Tables 3 and 4). Consequently, our study demonstrates that gender differences in IT-related behavior, in particular in trust situations, are related to neurobiology. Our study, therefore, corroborates the notions that "all human behavior that varies among individuals is partially heritable and correlated with measurable aspects of brains" (Turkheimer 1998, p. 782) and, moreover, that it is evident that

[17] However, abnormality on eBay does not exist only in cases of missing or very spare product descriptions. Rather, a very long, extremely detailed description, as well, is expected to result in perception of abnormality. Future research could test this hypothesis.

> there are sex influences at all levels of the nervous system....The picture of brain organization that emerges is of two complex mosaics—one male and one female—that are similar in many respects but very different in others. The way that information is processed though the two mosaics, and the behaviors that each produce, could be identical or strikingly different, depending on a host of parameters (Cahill 2006, p. 7).

The results of our investigation clearly support this concept.

5. In the present study, our theorizing draws upon the *empathizing–systemizing theory* (Baron-Cohen et al. 2005). Originally, this theory was developed to explain gender differences in autism. In essence, the theory posits that autism represents an extreme of the male information processing pattern (i.e., impaired empathizing and enhanced systemizing). Our study confirms the empathizing–systemizing theory in a trustworthiness evaluation task, because we found that women process Internet offers more emotionally, whereas men do so more cognitively. Hence, we demonstrate that neuroscience theories with original medical implications can have considerable implications for IS research and practice, too. In other words, the present study extends the scope of validity of the empathizing–systemizing theory (from medicine to online trust). This finding is relevant to both IS scholars and physicians.

As is common in research articles, the present study has limitations that should be considered. First, the interpretation of our empirical findings is based on a binary trustworthiness rating task in a controlled laboratory environment. Future studies could use more fine-grained scales, rather than binary ones, to evaluate the trustworthiness of Internet offers. Second, during the experiment, participants were required to lie still and were restrained with pads to prevent small motions during measurement sessions. The fMRI experiment, moreover, required participants to lie supine, within an acoustically noisy scanner. Obviously, such a situation is artificial, because in real life most online shoppers sit in front of their computers in a more or less comfortable and calm environment. Third, the present study investigated one particular product only (USB flash drives) in one specific online-brand frame (eBay). Furthermore, we used only one textual formulation for each component of Toulmin's model of argumentation. The use of different products and text formulations might trigger different brain activations, thereby resulting in different findings.

Despite these limitations, we believe that the present study contributes to a better understanding of the factors influencing perceived trustworthiness of Internet offers and, therefore, of online shopping trust. We now know that neurobiological factors are associated with variance in trustworthiness decisions in online settings, and as a result we can answer the question of *why* there are gender differences in online trust. Such differences exist because women and men recruit different brain areas when facing uncertainty, processing information, and deciding which offer to trust. Consequently, we can conclude that women and men are not only different in observable biological ways, but they are distinct on a deeper neural level as well—which until now has not been investigated in research on the subject of online trust.

The question of why women and men are different on this neural level remains open, however. It is possible, on the one hand, that biological factors such as genes (e.g., Cesarini et al. 2008) and hormones (e.g., Kosfeld et al. 2005) cause these gender differences. It is a well-established fact, for example, that gender differences exist concerning the production of oxytocin, considered to be the "trust hormone" (McLean and Anderson 2009). On the other hand, it is possible that socialization and experience, from both childhood and adulthood, are determinants of brain activation differences, because learning processes affect brain development via brain plasticity (e.g., Hyde 2007).[18] One intriguing study (Maguire et al. 2000) exemplifying this concept found that for cab drivers the size of the hippocampus (the brain region that plays a major role in the recall of complex routes) varied as a consequence of years spent driving taxis. Hence, learning and experience may change brain structure, even in adults. Altogether, we surmise that trust is influenced by both biological and socialization factors, and it will be rewarding to see what insight future research will reveal.

Acknowledgments

We would like to thank the senior editor, associate editor, and three anonymous reviewers for their excellent work in providing guidance on ways to improve the paper. We appreciate the generous support of Schindler Parent Meersburg by advancing the Schindler Parent Distinguished Guest Lecturer for Marketing. Finally, we thank Deborah C. Nester for proof-reading.

[18] Socialization and stereotypes have already been discussed in the IS literature as theoretical explanations for the existence of gender-specific differences in IT-related behavior (e.g., Adam et al. 2004; Lockheed 1985; Trauth 2002; Trauth et al. 2008; Truman and Baroudi 1994; Whitley 1997).

References

Adam, A., Howcroft, D., and Richardson, H. 2004. "A Decade of Neglect: Reflecting on Gender and IS," *New Technology, Work and Employment* (19:3), pp. 222-240.

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., and Damasio, A. R. 2005. "A Mechanism for Impaired Fear Recognition After Amygdala Damage," *Nature* (433:7021), pp. 68-72.

Adolphs, R., Tranel, D., and Damasio, A. R. 1998. "The Human Amygdala in Social Judgment," *Nature* (393:6684), pp. 470-474.

Alesina, A., and La Ferrara, E. 2002. "Who Trusts Others?," *Journal of Public Economic* (85:2), pp. 207-234.

Alkire, M. T., Haier, R. J., Fallon, J. H., and Cahill, L. 1998. "Hippocampal, but Not Amygdala, Activity at Encoding Correlates with Long-Term, Free Recall of Nonemotional Information," *Proceedings of the National Academy of Sciences* (95:24), pp. 14506-14510.

Allen, D. K., Colligan, D., Finnie, A., and Kern, T. 2000. "Trust, Power and Interorganizational Information Systems: The Case of the Electronic Trading Community TransLease," *Information Systems Journal* (10:1), pp. 21-40.

Ashburner, J., Neelin, P., Collins, L. D., Evans, A., and Friston, K. J. 1997. "Incorporating Prior Knowledge into Image Registration," *Neuroimage* (6:4), pp. 344-352.

Awad, N. F., and Ragowsky, A. 2008. "Establishing Trust in Electronic Commerce Through Online Word of Mouth: An Examination Across Genders," *Journal of Management Information Systems* (24:4), pp. 101-121.

Ba, S. L., and Pavlou, P. A. 2002. "Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior," *MIS Quarterly* (26:3), pp. 243-268.

Baier, A. 1986. "Trust and Antitrust," *Ethics* (96:2), pp. 231-260.

Bannerman, D. M., Rawlins, J. N. P., McHugh, S. B., Deacon, R. M. J., Yee, B. K., Bast, T., Zhang, W.-N., Pothuizen, H. H. J., and Feldon, J. 2004. "Regional Dissociations Within the Hippocampus: Memory and Anxiety," *Neuroscience & Biobehavioral Reviews* (28:3), pp. 273-283.

Baron-Cohen, S., Knickmeyer, R. C., and Belmonte, M. K. 2005. "Sex Differences in the Brain: Implications for Explaining Autism," *Science* (310:5749), pp. 819-823.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., and Fehr, E. 2008. "Oxytocin Shapes the Neural Circuitry of Trust and Trust Adaptation in Humans," *Neuron* (58:4), pp. 639-650.

Bechara, A., and Damasio, A. R. 2005. "The Somatic Marker Hypothesis: A Neural Theory of Economic Decision," *Games and Economic Behavior* (52:2), pp. 336-372.

Bechara, A., Damasio, H., Damasio, A. R., and Lee, G. P. 1999. "Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision Making," *Journal of Neuroscience* (19:13), pp. 5473-5481.

Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. 1997. "Deciding Advantageously Before Knowing the Advantageous Strategy," *Science* (275:5304), pp. 1293-1295.

Bechara, A., Tranel, D., and Damasio, H. 2000. "Characterization of the Decision-Making Deficit of Patients with Ventromedial Prefrontal Cortex Lesions," *Brain* (123:11), pp. 2189-2202.

Bechara, A., Tranel, D., Damasio, H., and Damasio, A. R. 1996. "Failure to Respond Autonomically to Anticipated Future Outcomes Following Damage to Prefrontal Cortex," *Cerebral Cortex* (6:2), pp. 215-225.

Bell, E., Willson, M. C., Wilman, A. H., Dave, S., and Silverstone, P. H. 2006. "Males and Females Differ in Brain Activation During Cognitive Tasks," *Neuroimage* (30:2), pp. 529-538.

Berg, J., Dickhaut, J., and McCabe, K. 1995. "Trust, Reciprocity and Social History," *Games and Economic Behavior* (10:1), pp. 122-142.

Bettman, J. R. 1973. "Perceived Risk and its Components: Model and Empirical Test," *Journal of Marketing Research* (10:2), pp. 184-190.

Blau, P. M. 1964. *Exchange and Power in Social Life*, New York: John Wiley & Sons.

Bolton, G. E., Katok, E., Ockenfels, A. 2004. "Trust among Internet Traders: A Behavioral Economics Approach," Working Paper Number 5, University of Cologne, Department of Economics.

Botvinick, M. M., Cohen, J. D., and Carter, C. S. 2004. "Conflict Monitoring and Anterior Cingulate Cortex: An Update," *Trends in Cognitive Sciences* (8:12), pp. 539-546.

Botvinick, M. M., Nystrom, L. E., Fissell, K., Carter, C. S., and Cohen, J. D. 1999. "Conflict Monitoring Versus Selection-for-Action in Anterior Cingulate Cortex," *Nature* (402:6758), pp. 179-181.

Broca, P. 1861. "Sur le volume et la forme du cerveau suivant les individus et suivant les races," *Bulletins et mimoires de la Societe d'Anthropologie de Paris* (2), pp. 139-207, 301-321, 441-446.

Broos, A. M. A. 2005. "Gender and Information and Communication Technologies (ICT) Anxiety: Male Self-Assurance and Female Hesitation," *CyberPsychology & Behavior* (8:1), pp. 21-31.

Brosig, J., Weimann, J., and Ockenfels, A. 2003. "The Effect of Communication Media on Cooperation," *German Economic Review* (4:2), pp. 217-241.

Buchan, N. R., Croson, R. T. A., and Solnick, S. 2008. "Trust and Gender: An Examination of Behavior and Beliefs in the Investment Game," *Journal of Economic Behavior & Organization* (68:3-4), pp. 466-476.

Bush, G., Luu, P., and Posner, M. I. 2000. "Cognitive and Emotional Influences in Anterior Cingulate Cortex," *Trends in Cognitive Sciences* (4:6), pp. 215-222.

Byrnes, J. P., Miller, D. C., and Schafer, W. D. 1999. "Gender Differences in Risk Taking: A Meta-Analysis," *Psychological Bulletin* (125:3), pp. 367-383.

Cahill, L. 2006. "Why Sex Matters for Neuroscience," *Nature Reviews Neuroscience* (7:6), pp. 477-484.

Campbell, C. 2000. "Shopaholics, Spendaholics, and the Question of Gender," in *I Shop, Therefore I Am: Compulsive Buying and the Search for Self* , A. Benson (ed.), NewYork: Aronson, pp. 57-75.

Canli, T., Desmond, J. E., Zhao, Z., and Gabrieli, J. D. E. 2002. "Sex Differences in the Neural Basis of Emotional Memories," *Proceedings of the National Academy of Sciences* (99:16), pp. 10789-10794.

Carter, C. S., Braver, T. S., Barch, D. M., Motvinick, M. M., Noll, D., and Cohen, J. D. 1998. "Anterior Cingulate Cortex, Error Detection, and the Online Monitoring of Performance," *Science* (280:5364), pp. 747-749.

Carter, L., and Belanger, F. 2005. "The Utilization of e-Government Services: Citizen Trust, Innovation and Acceptance Factors," *Information Systems Journal* (15:1), pp. 5-25.

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., and Wallace, B. 2008. "Heritability of Cooperative Behavior in the Trust Game," *Proceedings of the National Academy of Sciences* (105:10), pp. 3721-3726.

Cho, J. 2006. "The Mechanism of Trust and Distrust Formation and Their Relational Outcomes," *Journal of Retailing* (82:1), pp. 25-35.

Clark, L., Bechara, A., Damasio, H., Aitken, M. R. F., Sahakian, B. J., and Robbins, T. W. 2008. "Differential Effects of Insular and Ventromedial Prefrontal Cortex Lesions on Risky Decision Making," *Brain* (131:5), pp. 1311-1322.

Cook, K. S., Hardin, R., and Levi, M. 2005. *Cooperation Without Trust?*, New York: Russell Sage Publications.

Cosgrove, K. P., Mazure, C. M., and Staley, J. K. 2007. "Evolving Knowledge of Sex Differences in Brain Structure, Function, and Chemistry," *Biological Psychiatry* (62:8), pp. 847-855.

Critchley, H. D., Mathias, C. J., and Dolan, R. J. 2001. "Neural Activity in the Human Brain Relating to Uncertainty and Arousal During Anticipation," *Neuron* (29:2), pp. 537-545.

Crone, E. A., and van der Molen, M. W. 2004. "Developmental Changes in Real Life Decision Making: Performance on a Gambling Task Previously Shown to Depend on the Ventromedial Prefrontal Cortex," *Developmental Neuropsychology* (25:3), pp. 251-279.

Croson, R., and Buchan, N., 1999. "Gender and Culture: International Experimental Evidence from Trust Games," *American Economic Review* (89:2), pp. 386-391.

Cyr, D., and Bonanni, C. 2005. "Gender and Website Design in E-Business," *International Journal of Electronic Business* (3:6), pp. 565-582.

Cyr, D., Head, M., Larios, H., and Pan, B. 2009. "Exploring Human Images in Website Design: A Multi-Method Approach," *MIS Quarterly* (33:3), pp. 539-566.

Damasio, A. R. 1996. "The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex," *Philosophical Transactions of the Royal Society–B Biological Sciences* (351), pp. 1413-1420.

Darwin, C. 1871. *The Descent of Man*, London: Murray.

Delgado, M. R., Frank, R. H., and Phelps, E. A. 2005. "Perceptions of Moral Character Modulate the Neural Systems of Reward During the Trust Game," *Nature Neuroscience* (8:11), pp. 1611-1618.

Deppe, M., Schwindt, W., Kugel, H., Plassmann, H., and Kenning, P. 2005. "Nonlinear Responses Within the Medial Prefrontal Cortex Reveal When Specific Implicit Information Influences Economic Decision Making," *Journal of Neuroimaging* (15:2), pp. 171-182.

Devinsky, O., Morrell, M. J., and Vogt, B. A. 1995. "Contributions of Anterior Cingulate Cortex to Behavior," *Brain* (118:1), pp. 279-306.

Dimoka, A., and Davis, F. D. 2008. "Where Does TAM Reside in the Brain? The Neural Mechanisms Underlying Technology Adoption," in *Proceedings of the 29th International Conference on Information Systems*, Paris, December 14-17.

Dimoka, A., Pavlou, P. A., and Davis, F. D. 2007. "NEURO-IS: The Potential of Cognitive Neuroscience for Information Systems Research," in *Proceedings of the 28th International Conference on Information Systems*, Montreal, December 9-12.

Dittmar, H., Beattie, J., and Friese, S. 1996. "Objects, Decision Considerations, and Self-Image in Men's and Women's Impulse Purchases," *Acta Psychologica* (93:1-3), pp. 187-206.

Dittmar, H., and Drury, J. 2000. "Self-Image—Is it in the Bag? A Qualitative Comparison between 'Ordinary' and 'Excessive' Consumers," *Journal of Economic Psychology* (21:2), pp. 109-142.

Dittmar, H., Long, K., and Meek, R. 2004. "Buying on the Internet: Gender Differences in On-Line and Conventional Buying Motivations," *Sex Roles* (50:5-6), pp. 423-444.

Dollar, D., Fisman, R., and Gatti, R. 2001. "Are Women Really the 'Fairer' Sex? Corruption and Women in Government," *Journal of Economic Behavior and Organization* (46:4), pp. 423-429.

Eiseman, L. 2006. *Color: Messages and Meanings*, Gloucester: Hand Books Press.

Engell, A. D., Haxby, J. V., and Todorov, A. 2007. "Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala," *Journal of Cognitive Neuroscience* (19:9), pp. 1508-1519.

Ernst, M., Bolla, K., Mouratidis, M., Contoreggi, C., Matochik, J. A., Kurian, V., Cadet, J. L., Kimes, A. S., and London, E. D. 2002. "Decision Making in a Risk-Taking Task: A PET Study," *Neuropsychopharmacology* (26:5), pp. 682-691.

Fehr, E., and Camerer, C. F. 2007. "Social Neuroeconomics: The Neural Circuitry of Social Preferences," *Trends in Cognitive Sciences* (11:10), pp. 419-427.

Fellows, L. K., and Farah, M. J. 2007. "The Role of Ventromedial Prefrontal Cortex in Decision Making: Judgment Under Uncertainty or Judgment Per Se?," *Cerebral Cortex* (17:11), pp. 2669-2674.

Fogel, J., and Nehmad, E. 2009. "Internet Social Network Communities: Risk Taking, Trust, and Privacy Concerns," *Computers in Human Behavior* (25:1), pp. 153-160.

Frenkel, K. A. 1990. "Women and Computing," *Communications of the ACM* (33:11), pp. 34-46.

Friston, K. J. 1996. "Statistical Parametric Mapping and Other Analyses of Functional Imaging Data," in *Brain Mapping: The Methods*, A. W. Toga and J. C. Mazziotta (eds.), New York: Academic Press, pp. 363-386.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-B., Frith, C. D., and Frackowiak, R. S. J. 1995. "Statistical Parametric Maps in Functional Imaging: A General Linear Approach," *Human Brain Mapping* (2:4), pp. 189-210.

Gallivan, M. J. 2001. "Striking a Balance Between Trust Anti Control in a Virtual Organization: A Content Analysis of Open Source Software Case Studies," *Information Systems Journal* (11:4), pp. 277-304.

Gallivan, M. J., and Depledge, G. 2003. "Trust, Control and the Role of Interorganizational Systems in Electronic Partnerships," *Information Systems Journal* (13:2), pp. 159-190.

Gambetta, D. (ed.). 1988. *Trust: Making and Breaking Co-operative Relations*, Oxford: Basil Blackwell.

Garbarino, E., and Strahilevitz, M. 2004. "Gender Differences in the Perveived Risk of Buying Online and the Effects of Receiving a Site Recommendation," *Journal of Business Research* (57:7), pp. 768-775.

Gefen, D., Benbasat, I., and Pavlou, P. A. 2008. "A Research Agenda for Trust in Online Environments," *Journal of Management Information Systems* (24:4), pp. 275-286.

Gefen, D., Karahanna, E., and Straub, D. W. 2003. "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly* (27:1), pp. 51-90.

Gefen, D., and Ridings, C. M. 2005. "If You Spoke as She Does, Sir, Instead of the Way You Do: A Sociolinguistics Perspective of Gender Differences in Virtual Communities," *The DATA BASE for Advances in Information Systems* (36:2), pp. 78-92.

Gefen, D., and Straub, D. W. 1997. "Gender Differences in the Perception and Use of E-Mail: An Extension to the Technology Acceptance Model," *MIS Quarterly* (21:4), pp. 389-400.

Ghose, A., Ipeirotis, P. G., and Sundararajan, A. 2005. "Reputation Premiums in Electronic Peer-to-Peer Markets: Analyzing Textual Feedback and Network Structure," in *Proceedings of the 2005 ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems*, Philadelpiha, PA, August 22, pp. 150-154.

Glaeser, E. L, Laibson, D. I., Scheinkman, J. A., and Soutter, C. L. 2000. "Measuring Trust," *Quarterly Journal of Economics* (115:3), pp. 811-846.

Gläscher, J., Hampton, A. N., and O'Doherty, J. P. 2009. "Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making," *Cerebral Cortex* (19:2), pp. 483-495.

Gregg, D. G., and Scott, J. D. 2008. "A Typology of Complaints About eBay Sellers," *Communications of the ACM* (51:4), pp. 69-74.

Grimsley, M, and Meehan, A. 2007. "e-Government Information Systems: Evaluation-Led Design for Public Value and Client Trust," *European Journal of Information Systems* (16:2), pp. 134-148.

Hagoort, P., Indefrey, P., Brown, C., Herzog, H., Steinmetz, H., and Seitz, R. J. 1999. "The Neural Aircuitry Involved in the Reading of German Words and Pseudowords: A PET Study," *Journal of Cognitive Neuroscience* (11:4), pp. 383-398.

Haier, R. J., Jung, R. E., Yeo, R., Head, K., and Alkire, M. T. 2005. "The Neuroanatomy of General Intelligence: Sex Matters," *Neuroimage* (25:1), pp. 320-327.

Hall, C. C., Ariss, L., and Todorov, A. 2007. "The Illusion of Knowledge: When More Information Reduces Accuracy and Increases Confidence," *Organizational Behavior and Human Decision Processes* (103:2), pp. 277-290.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., and Gernsbacher, M. A. 2007. "The Science of Sex Differences in Science and Mathematics," *Psychological Science in the Public Interest* (8:1), pp. 1-51.

Heekeren, H. R., Marrett, S., Ruff, D. A., Bandettini, P. A., and Ungerleider, L. G. 2006. "Involvement of Human Left Dorsolateral Prefrontal Cortex in Perceptual Decision Making Is Independent of Response Modality," *Proceedings of the National Academy of Sciences* (103:26), pp. 10023-10028.

Hoeft, F., Watson, C. L., Kesler, S. R., Bettinger, K. E., and Reiss, A. L. 2008. "Gender Differences in the Mesocorticolimibic System During Computer Game-Play," *Journal of Psychiatric Research* (42:4), pp. 253-258.

Hoffman, D. L., Novak, T. P., and Peralta, M. 1999. "Building Consumer Trust Online," *Communications of the ACM* (42:4), pp. 80-85.

Horst, M., Kuttschreuter, M., and Gutteling, J. M. 2007. "Perceived Usefulness, Personal Experiences, Risk Perception and Trust as Determinants of Adoption of e-Government Services in the Netherlands," *Computers in Human Behavior* (23:4), pp. 1838-1852.

Huettel, S. A., Song, A. W., and McCarthy, G. 2009. *Functional Magnetic Resonance Imaging* (2nd ed.), Sunderland, MA: Sinauer Associates.

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., and Platt, M. L. 2006. "Neural Signatures of Economic Preferences for Risk and Ambiguity," *Neuron* (49:5), pp. 765-775.

Hyde, J. S. 2007. "New Directions in the Study of Gender Similarities and Differences," *Current Directions in Psychological Science* (16:5), pp. 259-263.

Jackson, L. A., Ervin, K. S., Gardner, P. D., and Schmitt, N. 2001. "Gender and the Internet: Women Communicating and Men Searching," *Sex Roles* (44:5-6), pp. 363-379.

Jarvenpaa, S. L., Shaw, T. R., and Staples, D. S. 2004. "Toward Contextualized Theories of Trust: The Role of Trust in Global Virtual Teams," *Information Systems Research* (15:3), pp. 250-267.

Johnson, T. J., and Kaye, B. K. 2009. "In Blog We Trust? Deciphering Credibility of Components of the Internet Among Politically Interested Internet Users," *Computers in Human Behavior* (25:1), pp. 175-182.

Junglas, I. A., Johnson, N. A., Steel, D. J., Abraham, C., Loghlin, P. M. 2007. "Identity Formation, Learning Styles and Trust in Virtual Worlds," *The DATA BASE for Advances in Information Systems* (38:4), pp. 90-96.

Kanwisher, N., and Yovel, G. 2006. "The Fusiform Face Area: A Cortical Region Specialized for the Perception of Faces," *Philosophical Transactions of the Royal Society–B Biological Sciences* (361), pp. 2109-2128.

Kenning, P. 2008. "The Influence of General Trust and Specific Trust on Buying Behavior," *International Journal of Retail & Distribution Management* (36:6), pp. 461-476.

Kenning, P., and Plassmann, H. 2005. "NeuroEconomics: An Overview from an Economic Perspective," *Brain Research Bulletin* (67:5), pp. 343-354.

Kenning, P., Deppe, M., Schwindt, W., Kugel, H., Plassmann, H. 2009. "The Good, the Bad and the Forgotten: An fMRI Study on Ad Liking and Ad Memory," *Advances in Consumer Research* (36), pp. 4-7.

Kenning, P., Plassmann, H., and Ahlert, D. 2007. "Application of Functional Magnetic Resonance Imaging for Market Research," *Qualitative Market Research: An International Journal* (10:2), pp. 135-152.

Kim, D., and Benbasat, I. 2006. "The Effects of Trust-Assuring Arguments on Consumer Trust in Internet Stores: Application of Toulmin's Model of Argumentation," *Information Systems Research* (17:3), pp. 286-300.

Kim, D. J., Ferrin, D. L., Rao, H. R. 2009. "Trust and Satisfaction, Two Stepping Stones for Successful e-Commerce Relationships: A Longitudinal Exploration," *Information Systems Research* (20:2), pp. 237-257.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. 2005. "Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange," *Science* (308:5718), pp. 78-83.

Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., and Loewenstein, G. 2007. "Neural Predictors of Purchases," *Neuron* (53:1), pp. 147-156.

Komiak, S. Y. X., and Benbasat, I. 2006. "The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents," *MIS Quarterly* (30:4), pp. 941-960.

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., and Ono, T. 2001. "Retrospective and Prospective Coding for Predicted Reward in the Sensory Thalamus," *Nature* (412:6846), pp. 546-549.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., and Fehr, E. 2005. "Oxytocin Increases Trust in Humans," *Nature* (435:7042), pp. 673-676.

Krain, A. L., Wilson, A. M., Arbuckle, R., Castellanos, F. X., and Milham, M. P. 2006. "Distinct Neural Mechanisms of Risk and Ambiguity: A Meta-Analysis of Decision Making," *Neuroimage* (32:1), pp. 477-484.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., Heinecke, A., and Grafman, J. 2007. "Neural Correlates of Trust," *Proceedings of the National Academy of Sciences* (104:50), pp. 20084-20089.

Kucian, K., Loenneker, T., Dietrich, T., Martin, E., and von Aster, M. 2005. "Gender Differences in Brain Activation Patterns During Mental Rotation and Number Related Cognitive Tasks," *Psychology Science* (47:1), pp. 112-131.

Kugler, T., Connolly, T., and Kausel, E. E. 2009. "The Effect of Consequential Thinking on Trust Game Behavior," *Journal of Behavioral Decision Making* (22:2), pp. 101-119.

Kuhnen, C. M., and Knutson, B. 2005. "The Neural Basis of Financial Risk Taking," *Neuron* (47:5), pp. 763-770.

Lee, D., and Seo, H. 2007. "Mechanisms of Reinforcement Learning and Decision Making in the Primate Dorsolateral Prefrontal Cortex," *Annals of the New York Academy of Sciences* (1104:1), pp. 108-122.

Lewicki, R. J., McAllister, D. J., and Bies, R. J. 1998. "Trust and Distrust: New Relationships and Realities," *Academy of Management Review* (23:3), pp. 438-458.

Lewis, J. D., and Weigert, A. 1985. "Trust as a Social Reality," *Social Forces* (63:4), pp. 967-985.

Lockheed, M. E. 1985. "Women, Girls, and Computers: A First Look at the Evidence," *Sex Roles* (13:3-4), pp. 115-122.

Luhmann, N. 1979. *Trust and Power*. Chichester, UK: John Wiley & Sons.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Ashburner, C. D., Frackowiak, R. S. J., and Frith, C. D. 2000. "Navigation-Related Structural Change in the Hippoccampi of Taxi Drivers," *Proceedings of the National Academy of Sciences* (97:8), pp. 4398-4403.

Mak, A. K. Y., Hu, Z., Zhang, J. X. X., Xiao, Z., and Lee, T. M. C. 2009. "Sex-Related Differences in Neural Activity During Emotion Regulation," *Neuropsychologia* (47:13), pp. 2900-2908.

Maddock, R. J., Garrett, A. S., and Buonocore, M. H. 2003. "Posterior Cingulate Cortex Activation by Emotional Words: fMRI Evidence from a Valence Decision Task," *Human Brain Mapping* (18:1), pp. 30-41.

McClure, E. B., Monk, C. S., Nelson, E. E., Zarahn, E., Leibenluft, E., Bilder, R. M., Charney, D. S., Ernst, M., and Pine, D. S. 2004. "A Developmental Examination of Gender Differences in Brain Engagement During Evaluation of Threat," *Biological Psychiatry* (55:11), pp. 1047-1055.

McCoy, A. N., and Platt, M. L. 2005. "Risk-Sensitive Neurons in Macaque Posterior Cingulate Cortex," *Nature Neuroscience* (8:9), pp. 1220-1227.

McHugh, S. B., Deacon, R. M. J., Rawlins, J. N. P., and Bannerman, D. M. 2004. "Amygdala and Ventral Hippocampus Contribute Differentially to Mechanisms of Fear and Anxiety," *Behavioral Neuroscience* (118:1), pp. 63-78.

McKnight, D. H., and Chervany, N. L. 2001. "Trust and Distrust Definitions: One Bite at a Time," in *Trust in Cyper-Societies*, R. Falcone, M. Singh, and Y.-H. Tan (eds.), Berlin/Heidelberg: Springer, pp. 27-54.

McKnight, D. H., and Choudhury, V. 2006. "Distrust and Trust in B2C e-Commerce: Do They Differ? in *Proceedings of the 8th International Conference on Electronic Commerce: The New E-Commerce: Innovations for Conquering Current Barriers, Obstacles, and Limitations to Conducting Successful Business on the Internet*, Fredericton, New Brunswick, Canada, August 13-16, pp. 482-491.

McKnight, D. H., Choudhury, V., and Kacmar, C. 2002a. "Developing and Validating Trust Measures for e-Commerce: An Integrative Typology," *Information Systems Research* (13:3), pp. 334-359.

McKnight, D. H., Choudhury, V., and Kacmar, C. 2002b. "The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: A Trust Building Model," *Journal of Strategic Information Systems* (11:3-4), pp. 297-323.

McKnight, D. H., Cummings, L. L., and Chervany, N. L. 1998. "Initial Trust Formation in New Organizational Relationships," *Academy of Management Review* (23:3), pp. 473-490.

McKnight, D. H., Kacmar, C. J., and Choudhury, V. 2004. "Dispositional Trust and Distrust Distinctions in Predicting High- and Low-Risk Internet Expert Advice Site Perceptions," *e-Service Journal* (3:2), pp. 35-58.

McLean, C. P., and Anderson, E. R. 2009. "Brave Men and Timid Women? A Review of the Gender Differences in Fear and Anxiety," *Clinical Psychology Review* (29:6), pp. 496-505.

Mechelli, A., Humphreys, G. W., Mayall, K., Olson, A., and Price, C. J. 2000. "Differential Effects of Word Length and Visual Contrast in the Fusiform and Lingual Gyri During Reading," *Proceedings of the Royal Society of London–B Biological Sciences* (267:1455), pp. 1909-1913.

Meyers-Levy, J. 1994. "Gender Differences in Cortical Organization: Social and Biochemical Antecedents and Advertising Consequences," in *Attention, Attitude, and Affect in Response to Advertising*, E. M. Clark, T. C. Brock, and D. W. Stewart (eds.), Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 107-122.

Meyers-Levy, J., and Maheswaran, D. 1991. "Exploring Differences in Males' and Females' Processing Strategies," *Journal of Consumer Research* (18:1), pp. 63-70.

Meyers-Levy, J., and Sternthal, B. 1991. "Gender Differences in the Use of Message Cues and Judgments," *Journal of Marketing Research* (28:1), pp. 84-96.

Milad, M. R., Quirk, G. J., Pitman, R. K., Orr, S .P., Fischl, B., and Rauch, S .L. 2007. "A Role of the Human Dorsal Anterior Cingulate Cortex in Fear Expression," *Biological Psychiatry* (62:10), pp. 1191-1194.

Morgan, R. M, and Hunt, S .D. 1994. "The Commitment–Trust Theory of Relationship Marketing," *Journal of Marketing* (58:3), pp. 20-38.

Nicolaou, A. I., and McKnight, D. H. 2006. "Perceived Information Quality in Data Exchanges: Effects on Risk, Trust, and Intention to Use," *Information Systems Research* (17:4), pp. 392-414.

O'Doherty, J. 2004. "Reward Representations and Reward-Related Learning in the Human Brain: Insights from Neuroimaging," *Current Opinion in Neurobiology* (14:6), pp. 769-776.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R. J. 2004. "Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning," *Science* (304:5669), pp. 452-454.

Ou, C. X. J., and Sia, C. L. 2009. "To Trust or to Distrust, That Is the Question: Investigating the Trust–Distrust Paradox," *Communications of the ACM* (52:5), pp. 135-139.

Oxedine, A., Borgida, E., Sullivan, J. L., and Jackson, M. S. 2003. "The Importance of Trust and Community in Developing and Maintaining a Community Electronic Network," *International Journal of Human-Computer Studies* (58:6), pp. 671-696.

Paul, D. L., and McDaniel, R .R. 2004. "A Field Study of the Effect of Interpersonal Trust on Virtual Collaborative Relationship Performance," *MIS Quarterly* (28:2), pp. 183-227.

Paulus, M. P., and Frank, L. R. 2003. "Ventromedial Prefrontal Cortex Activation Is Critical for Preference Judgments," *Neuroreport* (14:10), pp. 1311-1315.

Paulus, M. P., Rogalsky, C., Simmons, A., Feinstein, J. S., and Stein, M. B. 2003. "Increased Activation in the Right Insula During Risk-Taking Decision Making Is Related to Harm Avoidance and Neuroticism," *Neuroimage* (19:4), pp. 1439-1448.

Pavlou, P. A. 2003. "Consumer Acceptance of Electronic Commerce—Integrating Trust and Risk, with the Technology Acceptance Model," *International Journal of Electronic Commerce* (7:3), pp. 69-103.

Pavlou, P. A., and Dimoka, A. 2006. "The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Information Systems Research* (17:4), pp. 392-414.

Pavlou, P. A., and Gefen, D. 2004. "Building Effective Online Marketplaces with Institution-Based Trust," *Information Systems Research* (15:1), pp. 37-59.

Pavlou, P. A., Liang, H., and Xue, Y. 2007. "Understanding and Mitigating Uncertainty in Online Exchange Relationships: A Principal–Agent Perspective," *MIS Quarterly* (31:1), pp. 105-136.

Phillips, D. M., and Stanton, J. L. 2004. "Age-Related Differences in Advertising: Recall and Persuasion," *Journal of Targeting, Measurement & Analysis for Marketing* (13:1), pp. 7-20.

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., Williams, S. C. R., Bullmore, E. T., Brammer, M., and Gray, J. A. 1998. "Neural Responses to Facial and Vocal Expressions of Fear and Disgust," *Proceedings of the Royal Society of London–B Biological Sciences* (265:1408), pp. 1809-1817.

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., Bullmore, E. T., Perrett, D. I., Rowland, D., Williams, S. C. R., Gray, J. A., and David, A. S. 1997. "A Specific Neural Substrate for Perceiving Facial Expressions of Disgust," *Nature* (389:6650), pp. 495-498.

Piccoli, G., and Ives, B. 2003. "Trust and the Unintended Effexts of Behavior Control in Virtual Teams," *MIS Quarterly* (27:3), pp. 365-395.

Ploghaus, A., Tracey, I., Gati, J. S., Clare, S., Menon, R. S., Matthews, P. M., and Rawlins, J. N. P. 1999. "Dissociating Pain from its Anticipation in the Human Brain," *Science* (284:5422), pp. 1979-1981.

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., and Nichols, T. E. 2007. "Guidelines for Reporting an fMRI Study," *Neuroimage* (40:2), pp. 409-414.

Powell, P. L., and Johnson, J. E. V. 1995. "Gender and DSS Design: The Research Implications," *Decision Support Systems* (14:1), pp. 27-58.

Preuschoff, K., Bossaerts, P., and Quartz, S. R. 2006. "Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures," *Neuron* (51:3), pp. 381-390.

Preuschoff, K., Quartz, S. R., and Bossaerts, P. 2008. "Human Insula Activation Reflects Risk Prediction Errors as Well as Risk," *Journal of Neuroscience* (28:11), pp. 2745-2752.

Price, C. J., Moore, C. J., and Frackowiak, R. S. J. 1996. "The Effect of Varying Stimulus Rate and Duration on Brain Activity During Reading," *Neuroimage* (3:1), pp. 40-52.

Qiu, L. Y., and Benbasat, I. 2005. "Online Consumer Trust and Live Help Interfaces: The Effects of Text-to-Speech Voice and Three-Dimensional Avatars," *International Journal of Human-Computer Interaction* (19:1), pp. 75-94.

Ravaja, N., Kallinen, K., Saari, T., and Keltikangas-Jarvinen, L. 2004. "Suboptimal Exposure to Facial Expressions When Viewing Video Messages from a Small Screen: Effects on Emotion, Attention, and Memory," *Journal of Experimental Psychology: Applied* (10:2), pp. 120-131.

Resnick, P., and Zeckhauser, R. 2002. "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System," in *The Economics of the Internet and E-Commerce (Advances in Applied Microeconomics, Volume 11)*, M. R. Baye (ed.), Amsterdam: Elsevier Science, pp. 127-157.

Rodgers, S., and Harris, M. A. 2003. "Gender and e-Commerce: An Exploratory Study," *Journal of Advertising Research* (43:3), pp. 322-329.

Rotter, J. B. 1967. "A New Scale for the Measurement of Interpersonal trust," *Journal of Personality* (35:4), pp. 651-655.

Rushton, J. P. 1992. "Cranial Capacity Related to Sex, Rank, and Race in a Stratified Random Sample of 6,325 U.S. Military Personnel," *Intelligence* (16:3-4), 401-413.

Rushton, J. P., and Ankney, C. D. 1996. "Brain Size and Cognitive Ability: Correlations with Age, Sex, Social Class, and Race," *Psychonomic Bulletin & Review* (3:1), pp. 21-36.

Salloum, J. B., Saad, Z., Bodurka, J., Rio, D., Rawlings, R., and Hommer, D. W. 2008. "Gender Differences in Fusiform Gyrus Response Relative to Arousal Condition," *Biological Psychiatry* (63:7), p. 61.

Samanez-Larkin, G. R., Hollon, N. G., Carstensen, L. L., and Knutson, B. 2008. "Individual Differences in Insular Sensitivity During Loss Anticipation Predict Avoidance Learning," *Psychological Science* (19:4), pp. 320-323.

Sánchez-Franco, M. J. 2006. "Exploring the Influence of Gender on the Web Usage Via Partial Least Squares," *Behaviour & Information Technology* (25:1), pp. 19-36.

Sánchez-Franco, M. J., Ramos, A. F. V., and Velicia, F. A. M. 2009. "The Moderating Effect of Gender on Relationship Quality and Loyalty Toward Internet Service Providers," *Information & Management* (46:3), pp. 196-202.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game," *Science* (300:5626), pp. 1755-1758.

Savoy, R. L. 2005. "Experimental Design in Brain Activation MRI: Cautionary Tales," *Brain Research Bulletin* (67:5), pp. 361-367.

Schul, Y., Mayo, R., Burnstein, E. 2004. "Encoding Under Trust and Distrust: The Spontaneous Activation of Incongruent Cognitions," *Journal of Personality and Social Psychology* (86:5), pp. 668-679.

Schul, Y., Mayo, R., Burnstein, E. 2008. "The Value of Distrust," *Journal of Experimental Social Psychology* (44:5), pp. 1293-1302.

Schulte-Rüther, M., Markowitsch, H. J., Sha, N. J., Fink, G. R., and Piefke, M. 2008. "Gender Differences in Brain Networks Supporting Empathy," *Neuroimage* (42:1), pp. 393-403.

Schultz, W., Dayan, P., and Montague, P. R. 1997. "A Neural Substrate of Prediction and Reward," *Science* (275: 5306), pp. 1593-1599.

Schumacher, P., and Morahan-Martin, J. 2001. "Gender, Internet and Computer Attitudes and Experiences," *Computers in Human Behavior* (17:1), pp. 95-110.

Seo, H., and Lee, D. 2007. "Temporal Filtering of Reward Signals in the Dorsal Anterior Cingulate Cortex During a Mixed-Strategy Game," *Journal of Neuroscience* (27:31), pp. 8366-8377.

Seybert, H. 2007. "Gender Differences in the Use of Computers and the Internet," Eurostat: Statistics in Focus, Population and Social Conditions, 119/2007 (available online at http://www.ifap.ru/library//book243.pdf).

Shaub, M. K. 1996. "Trust and Suspicion: The Effects of Situational and Dispositional Factors on Auditors' Trust of Clients," *Behavioral Research in Accounting* (8:1), pp. 154-174.

Sheehan, K. B. 1999. "An Investigation of Gender Differences in On-Line Privacy Concerns and Resultant Behaviors," *Journal of Direct Marketing* (13:4), pp. 24-38.

Shirao, N., Okamoto, Y., Mantani, T., Okamoto, Y., and Yamawaki, S. 2005. "Gender Differences in Brain Activity Generated by Unpleasant Word Stimuli Concerning Body Image: An fMRI Study," *British Journal of Psychiatry* (186:1), pp. 48-53.

Simmons, A., Matthews, S. C., Stein, M. B., and Paulus, M. P. 2004. "Anticipation of Emotionally Aversive Visual Stimuli Activates Right Insula," *Neuroreport* (15:14), pp. 2261-2265.

Simon, S. J. 2001. "The Impact of Culture and Gender on Web Sites: An Empirical Study," *The DATA BASE for Advances in Information Systems* (32:1), pp. 18-37.

Snijders, C., and Keren, G. 1999. "Determinants of Trust," in *Games and Human Behavior*, D. Budescu, I. Erev, and R. Zwick (eds.), Mahwah, NJ: Lawrence Erlbaum Associates, pp. 355-386.

Starrfelt, R., and Gerlach, C. 2007. "The Visual What For Area: Words and Pictures in the Left Fusiform Gyrus," *Neuroimage* (35:1), pp. 334-342.

Steininger, K., Riedl, R., Roithmayr, F., and Mertens, P. 2009. "Fads and Trends in Business and Information Systems Engineering and Information Systems Rresearch: A Comparative Literature Analysis," *Business & Information Systems Engineering* (1:6), pp. 411-428.

Stenstrom, E., Stenstrom, P., Saad, G., and Cheikhrouhou, S. 2008. "Online Hunting and Gathering: An Evolutionary Perspective on Sex Differences in Website Preferences and Navigation," *IEEE Transactions on Professional Communication* (51:2), pp. 155-168.

Stoll, M., Baecke, S., and Kenning, P. 2008. "What They See Is What They Get? An fMRI-Study on Neural Correlates of Attractive Packaging," *Journal of Consumer Behavior* (7:4-5), pp. 342-359.

Sutter, M., and Kocher, M. G. 2007. "Trust and Trustworthiness Across Different Age Groups," *Games and Economic Behavior* (59:2), pp. 364-382.

Swaminathan, V., Lepkowska-White, E., and Rao, B. P. 1999. "Browsers or Buyers in Cyberspace? An Investigation of Factors Influencing Electronic Exchange," *Journal of Computer-Mediated Communication* (5:2) (http://jcmc.indiana.edu/vol5/issue2/swaminathan.htm).

Swamy, A., Knack, S., Lee, Y., and Azfar, O. 2001. "Gender and Corruption," *Journal of Development Economics* (64:1), pp. 25-55.

Teo, T. S. H., Srivastava, S. C., and Jiang, L. 2008. "Trust and Electronic Government Success: An Empirical Study," *Journal of Management Information Systems* (25:3), pp. 99-131.

Terrell, F., and Barrett, R. 1979. "Interpersonal Trust Among College Students as a Function of Race, Sex and Socioeconomic Class," *Perceptual and Motor Skills* (48:3), pp. 1194.

Todorov, A. 2008. "Evaluating Faces on Trustworthiness: An Extension of Systems for Recognition of Emotions Signaling Approach/Avoidance Behaviors," *Annals of the New York Academy of Sciences* (1124), pp. 208-224.

Todorov, A., Baron, S. G., and Oosterhof, N. N. 2008. "Evaluating Face Trustworthiness: A Model Based Approach," *Social Cognitive and Affective Neuroscience* (3:2), pp. 119-127.

Toulmin, S. 1958. *The Use of Argument*, Cambridge, UK: Cambridge University Press.

Trauth, E. M. 2002. "Odd Girl Out: An Individual Differences Perspective on Women in the IT Profession," *Information Technology & People* (15:2), pp. 98-118.

Trauth, E. M., Quesenberry, J. L., and Yeo, B. 2008. "Environmental Influences on Gender in the IT Workforce," *The DATA BASE for Advances in Information Systems* (39:1), pp. 8-32.

Treaster, J. 1999. "Equal Opportunity in Mexico City: Counting on Women to Be More Honest than Men," *New York Times*, August 15, Section 4, p. 3.

Tricomi, E. M., Delgado, M. R., and Fiez, J. A. 2004. "Modulation of Caudate Activity by Action Contingency," *Neuron* (41:2), pp. 281-292.

Truman, G. E., and Baroudi, J. J. 1994. "Gender Differences in the Information Systems Managerial Ranks: An Assessment of Potential Discriminatory Practices," *MIS Quarterly* (18:2), pp. 129-142.

Turkheimer, E. 1998. "Heritability and Biological Explanation," *Psychological Review* (105:4), pp. 782-791.

Van der Werf, Y. D., Tisserand, D. J., Visser, P. J., Hofman, P. A. M., Vuurman, E., Uylings, H. B. M., and Jolles, J. 2001. "Thalamic Volume Predicts Performance on Tests of Cognitive Speed and Decreases in Healthy Aging: A Magnetic Resonance Imaging-Based Volumetric Analysis," *Cognitive Brain Research* (11:3), pp. 377-385.

Van Slyke, C., Comunale, C. L., and Belanger, F. 2002. "Gender Differences in Perceptions of Web-Based Shopping," *Communications of the ACM* (45:7), pp. 82-86.

Van Welsum, D., and Montagnier, P. 2007. "ICTs and Gender," DSTI/ICCP/IE(2006)9/Final, Working Party on the Information Economy, Organisation for Economic Co-operation and Development, Paris (available online at http://www.oecd.org/dataoecd/16/33/38332121.pdf).

Van't Wout, M., Kahn, R. S., Sanfey, A. G., and Aleman, A. 2005. "Repetitive Transcranial Magnetic Stimulation Over the Right Dorsolateral Prefrontal Cortex Affects Strategic Decision Making," *Neuroreport* (16:7), pp. 1849-1852.

Vance, A., Elie-Dit-Cosaque, C., and Straub, D. W. 2008. "Examining Trust in Information Technology Artifacts," The Effects of System Quality and Culture," *Journal of Management Information Systems* (24:4), pp. 73-100.

Venkatesh, V., and Morris, M. G. 2000. "Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior," *MIS Quarterly* (24:1), pp. 115-139.

Wager, T. D., Phan, K. L., Liberzon, I., and Taylor, S. F. 2003. "Valence, Gender, and Lateralization of Functional Brain Anatomy in Emotion: A Meta-Analysis of Findings from Neuroimaging," *Neuroimage* (19:3), pp. 513-531.

Wang, W. Q., and Benbasat, I. 2008. "Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for e-Commerce," *Journal of Management Information Systems* (24:4), pp. 249-273.

Weiser, E. B. 2000. "Gender Differences in Internet Use Patterns and Internet Application Preferences: A Two-Sample Comparison," *CyperPsychology & Behavior* (3:2), pp. 167-178.

Welty, B., and Becerra-Fernandez, I. 2001. "Managing Trust and Commitment in Collaborative Supply Chain Relationships," *Communications of the ACM* (44:6), pp. 67-73.

Whitley, B. E. 1997. "Gender Differences in Computer-Related Attitudes and Behavior: A Meta-Analysis," *Computers in Human Behavior* (13:1), pp. 1-22.

Winston, J. S., Strange, B. A., O'Doherty, J., and Dolan, R. J. 2002. "Automatic and Intentional Brain Responses During Evaluation of Trustworthiness of Faces," *Nature Neuroscience* (5:3), pp. 277-283.

Wolin, L. D., and Korgaonkar, P. 2003. "Web Advertising: Gender Differences in Beliefs, Attitudes, and Behavior," *Internet Research: Electronic Networking Applications and Policy* (13:5), pp. 375-385.

Wright, T., and Sharp, E. 1979. "Content and Grammatical Sex Bias on the Interpersonal Trust Scale and Differential Trust Toward Women and Men," *Journal of Consulting and Clinical Psychology* (47:1), pp. 72-85.

Zak, P. J., and Knack, S. 2001. "Trust and Growth," *Economic Journal* (111:470), pp. 295-321.

Zak, P. J., Kurzban, R., and Matzner, W. T. 2005. "Oxytocin Is Associated with Human Trustworthiness," *Hormones and Behavior* (48:5), pp. 522-527.

Zucker, L. G. 1986. "Production of Trust: Institutional Sources of Economic Structure," in *Research in Organizational Behavior* (Volume 8), B. M. Staw and L. L. Cummings (eds.), Greenwich, CT: JAI Press, pp. 53-111.

About the Authors

René Riedl is an associate professor of Business Informatics at the University of Linz, Austria. He serves on the executive board of the Institute of Human Resources and Organizational Development in Management at the University of Linz. His main research interests are NeuroIS and IT management. He has published several IT-related books and his research has appeared in *Business & Information Systems Engineering*, *WIRTSCHAFTSINFORMATIK*, *Behavior Research Methods*, *NeuroPsychoEconomics,* and the Proceedings of the International Conference on Information Systems, among others. In 2006, René won a best paper award at the Conference of the Society for NeuroPsychoEconomics. In spring 2007, he was a guest researcher at Harvard University. He is the 2009 SchindlerParent Distinguished Guest Lecturer at Zeppelin University, Germany.

Marco Hubert is a Ph.D. student in Marketing at Zeppelin University, Germany. His overall research interest is consumer behavior, consumer neuroscience, and neuroeconomics. His work has been published in *Journal of Consumer Behaviour* and *International Journal of Retail and Distribution Management*. He has presented papers at a number of conferences including the Conference of the Society for Neuroeconomics.

Peter H. Kenning is a professor of Marketing at Zeppelin University, Germany. His overall research interest is consumer behavior, consumer neuroscience, neuroeconomics, and marketing management. His work has been published widely in outlets including *Management Decisions*, *Journal of Consumer Behaviour*, *Journal of Economic Psychology*, *Advances in Consumer Research*, as well as in the *Journal of Neuroimaging*, *Neuroreport*, and *Brain Research Bulletin*. He has presented papers at a number of conferences including the AMA Conference, the EMAC and ACR. He received several best-paper awards and grants from the German government.

Appendix

USB Flash Drive Images

The USB flash drive images used in the pretest and fMRI study are shown below. All USB flash drive images were selected from eBay and Amazon and they were all black, silver, anthracite, and grey (Eiseman 2006).

Journal of Neuroscience, Psychology, and Economics
2012, Vol. 5, No. 2, 63–91

1937-321X/11/$12.00 DOI: 10.1037/a0026318

The Biology of Trust: Integrating Evidence From Genetics, Endocrinology, and Functional Brain Imaging

René Riedl
University of Linz, Austria

Andrija Javor
Linz General Hospital, Austria

Trust is among the most important factors in human life, as it pervades almost all domains of society. Although behavioral research has revealed a number of insights into the nature of trust, as well as its antecedents and consequences, an increasing number of scholars have begun to investigate the topic from a biological perspective to gain a deeper understanding. These biological investigations into trust have been carried out on three levels of analysis: genes, endocrinology, and the brain. Based on these three levels, we present a review of the literature on the biology of trust. Moreover, we integrate our findings into a conceptual framework which unifies the three levels of analysis, and we also link the biological levels to trust behavior. The results show that trust behavior is at least moderately genetically predetermined. Moreover, trust behavior is associated with specific hormones, in particular oxytocin, as well as specific brain structures, which are located in the basal ganglia, limbic system, and the frontal cortex. Based on these results, we discuss both methodological and thematic implications.

Keywords: trust, biology, hormones, fMRI, oxytocin

Determining whom to trust and whom not to trust has been crucial since the early days of ancient civilizations. Trust in individuals who turned out not to be trustworthy resulted in death in many situations from the Stone Age to the Middle Ages. However, although in today's world the possible consequences of trusting untrustworthy people are not as directly related to survival as they were in former times, breached trust may result in severe negative feelings and/or economic consequences—imagine, for example, the spouse who is cheated in marriage or an online shopper whose product, although already paid for, is not delivered.

Economic research has shown that low trust, as the result of repeatedly breached trust, leads to a low rate of investments, which in turn impedes new businesses and employment (Zak & Fakhar, 2006). As a consequence, low trust countries are typically poor countries, and trust is, therefore, among the strongest predictors of poverty identified by economists (Zak & Knack, 2001). Overall, trust is among the most important factors in human life, as it pervades almost all domains of society, ranging from love, friendship, economic collaborations, medicine, and politics, to the steadily increasing number of human interactions in anonymous virtual Internet environments (Falk & Kosfeld, 2006; Füllbrunn, Richwien, & Sadrieh, 2011; Gefen, Benbasat, & Pavlou, 2008; Lee & Lin, 2009; Luhmann, 1979; Nguyen et al., 2009; Smith, 2010).

Against the background of the ubiquity of trust in human societies, both in traditional and virtual environments, a vast amount of research in several scientific disciplines has investigated the nature of trust, as well as its antecedents (e.g., facial expressions in personal interactions among humans or well-designed user interfaces in computer-mediated interactions) and consequences (e.g., information disclosure or cooper-

This article was published Online First November 28, 2011.
René Riedl, Department of Business Informatics–Information Engineering, University of Linz, Linz, Austria; Andrija Javor, Department of Neurology and Psychiatry, Linz General Hospital, Linz, Austria.
We thank Bernd Weber for his work in providing guidance on ways to improve this article. We also acknowledge the excellent comments during the review process. Finally, we are grateful to the participants of the Gmunden Retreat on NeuroIS 2011 (see www.NeuroIS.org) for their thoughtful comments on several parts of this review.
Correspondence concerning this article should be addressed to Dr. René Riedl, University of Linz, Department of Business Informatics–Information Engineering, Altenberger Strasse 69, 4040 Linz, Austria. E-mail: rene.riedl@jku.at

ative behavior). Psychologists, sociologists, economists, as well as management and information systems researchers, among scholars in other academic fields, have investigated trust, both theoretically and empirically (Gefen et al., 2008; Rousseau, Sitkin, Burt, & Camerer, 1998; Seppänen, Blomqvist, & Sundqvist, 2007; Swan, Bowers, & Richardson, 1999).

Despite the large variety of disciplines that have studied the phenomenon, there is a surprising consensus regarding the conceptualization of trust. Drawing upon the work of Fishbein and Ajzen (1975) and Coleman (1990), many scholars describe trust as a *behavior* which makes one party, the trustor, vulnerable to the actions of another party, the trustee. This behavior of a trustor is influenced by his or her *beliefs* about the trustee's trustworthiness, which in turn affect his or her *attitudes* toward the trustee and subsequent *behavioral intentions*. Moreover, trust behavior in a specific situation is also influenced by the trustor's general level of trust (i.e., trust disposition), as well as by an individual's risk preferences.

Important characteristics of a trustee are ability, benevolence, and integrity (Mayer, Davis, & Schoorman, 1995). If a trustor believes in the trustworthiness of a trustee, he or she believes that the trustee (a) has skills and competencies that are important for the relationship (ability), (b) means well toward the trustor aside from an egocentric profit motive (benevolence), and (c) adheres to a set of principles that the trustor finds acceptable (integrity). The importance of each characteristic may vary as a function of the inner states of a trustor (e.g., high risk perceptions regarding possible betrayal may turn one's attention to benevolence rather than ability) and context factors such as the importance of the trust decision for survival (e.g., in tandem skydiving, the ability of the instructor is expected to be more important than benevolence).

Figure 1 illustrates the structure of a trust situation. A person, the trustor, who meets an unknown individual, the trustee, for the first time, will form his or her beliefs about the individual's trustworthiness on the basis of stimuli (e.g., appearance, facial expression, or self-control of the trustee; Adolphs et al., 2005; Righetti & Finkenauer, 2011; Todorov, 2008; Winston, Strange, O'Doherty, & Dolan, 2002), which in turn affects the attitude toward the individual, subsequent behavioral intentions, and ultimately actual behavior (e.g., trust/distrust and approach/avoidance). Moreover, the trustor's disposition to trust, as well as his or her risk preferences, may moderate each linkage in this causal chain (e.g., Fehr, 2009a). (We

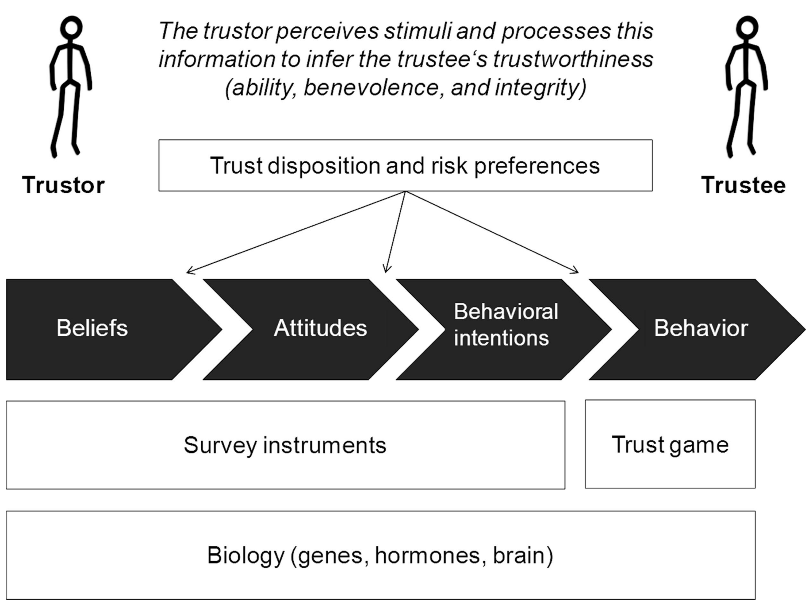


Figure 1. Structure of a trust situation.

describe the lower part in Figure 1 in the next section.)

Although behavioral trust research has revealed a number of insights into the nature of trust, as well as its antecedents and consequences (Rousseau et al., 1998; Seppänen et al., 2007; Swan et al., 1999), scholars have also started to investigate the topic from a biological perspective during the past decade. These biological investigations into trust can be classified into three groups, namely genetics (e.g., Cesarini et al., 2008), endocrinology (e.g., Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005), and brain functionality (e.g., King-Casas et al., 2005). One reason for this development toward biologically oriented trust research is the availability of powerful methods such as functional MRI (fMRI). A further reason is the insight that all human behavior that varies among individuals is associated, at least partly, with biological factors, in particular those related to the nervous system (e.g., Cacioppo, Berntson, Sheridan, & McClintock, 2000; Turkheimer, 1998). Hence, the consideration of biological factors is indispensible to explain a large proportion of the variance in human trust behavior (Fehr & Camerer, 2007).

In the present article, we review the literature that has investigated trust from a biological perspective. Extensive investigation revealed a large number of studies on this topic. So far, a vast amount of brain imaging studies, mostly using fMRI, has revealed several brain regions associated with trust. Moreover, endocrinological studies show that a number of hormones affect trust. Finally, recent gene-based research has demonstrated that at least a moderate degree of human trust behavior is genetically predetermined.

Due to the large number of biologically oriented studies on trust that we identified, a more complete understanding of the research findings would be possible if the studies were integrated into a unifying framework. Therefore, in the present article we discuss existing research on the biological foundations of trust within a conceptual framework that distinguishes three levels of analysis, namely genetic, hormonal, and at the brain level. Also, we discuss how these biological factors influence trust behavior.

The remainder of this article is structured as follows: In section 1, we highlight important foundations of trust research. First, we present a model which we use to structure our review. Second, we discuss the measurement of trust. In section 2, we outline the biology of trust along three levels of analysis (genes, hormones, and the brain), and we make explicit that the influence of the genetic and hormonal levels on trust behavior is mediated by the brain level (i.e., activation in specific brain areas). Then, in section 3, we integrate the findings of the three levels of analysis into a conceptual framework. Finally, in section 4, we summarize our findings, present implications for future research, and provide concluding comments.

1. Foundations of Trust

Biology and Trust Behavior: A Model

The question of how *nature* and *nurture* contribute to the manifestation of human behavior, such as trust behavior, has been one of the most fundamental research issues in a number of scientific disciplines, particularly in psychology. Though there have been extremists who have believed that either the biological influences of nature (e.g., genes) or the environmental influences of nurture (e.g., socialization) predominate, today most scientists agree that both are important and neither is deterministic (Johnson, 2007). Empirical evidence substantiates the notion that human behavior is the result of the complex interplay between both biological and environmental factors (e.g., Bouchard, 1994; Cacioppo et al., 2000).

Figure 2 illustrates the relations between behavior, biology, and environment. It is shown that human behavior is influenced by (a) biological factors (genes, hormones, and the brain) and (b) environmental factors (e.g., socialization, culture, experience, and task demands). While human behavior is directly observable (e.g., one person trusts another one and therefore exhibits approach behavior), biological factors are typically not (e.g., activation in brain regions associated with trust is not directly observable). Moreover, the interrelationship between biological and environmental factors is illustrated by the gray arrows in Figure 2.

In this article, we review the biologically oriented literature on trust. We discuss the relation between biological factors (genes, hormones, and the brain) and trust behavior. Moreover, we outline the relations among the three biological factors in the trust context. Research

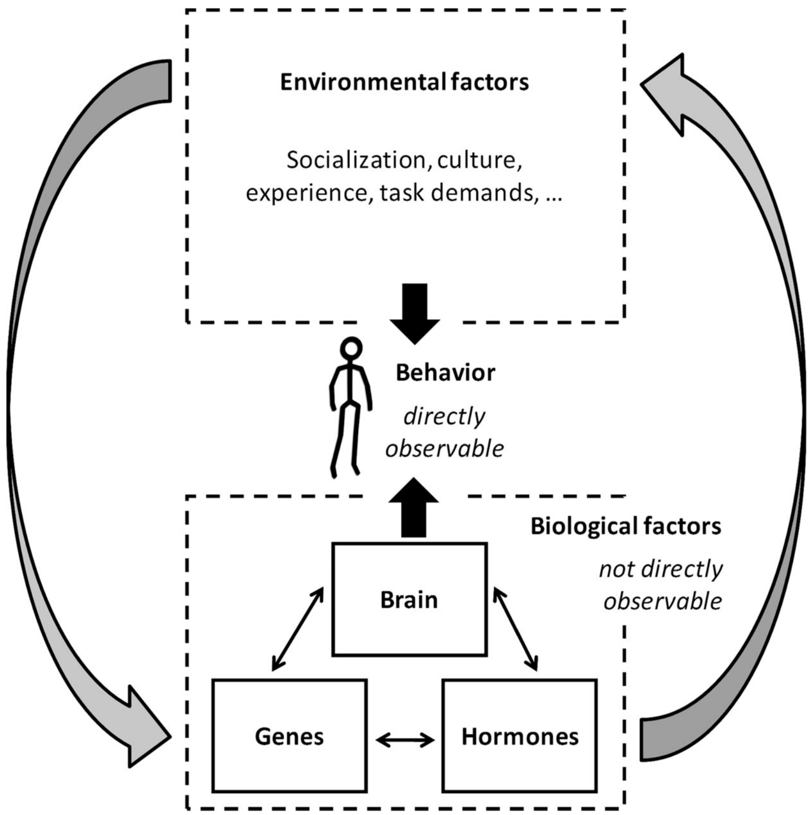


Figure 2. Model of behavior, biology, and environment.

(e.g., Cacioppo et al., 2000) indicates that genes, hormones, and the brain are interrelated components. For example, genes influence the production and release of hormones and the anatomy of the brain and its processing mechanisms. Moreover, hormones influence the activation of brain regions because certain regions have receptors for certain hormones. Finally, to state another example, the brain influences hormones because the brain regulates the production and release of hormones. The interrelationships among the biological factors are illustrated by the arrows in Figure 2. It is important to note that the influence of genes and hormones on behavior is mediated by activation in specific brain areas.

Note that in this article we do not discuss the direct influence of environmental factors on trust behavior (Zak & Fakhar, 2006). That is, we do not discuss topics such as the influence of socialization and culture on trust behavior. Such a discussion could be a fruitful topic for future studies (see, e.g., Bjornskov, 2007 and Welch et al., 2005).

The Measurement of Trust: Survey Instruments and The Trust Game

Biological research on trust is based on measurements that are often different from the conceptualizations used in the behavioral sciences (e.g., those used in psychology, economics, management science, and information systems research). Hence, to fully understand the biological foundations of trust, as well as their implications for trust behavior, it is important to outline the major conceptual differences.

Based on extensive reviews of the literature, research (McKnight & Chervany, 2001; Seppänen et al., 2007; Swan et al., 1999) has found that trust is often conceptualized as a belief, attitude, intention, or behavior (see Figure 1).

Such conceptualizations are used to develop survey instruments which make possible the measurement of trust. For example, if trust is conceptualized as a *belief*, the construct is operationalized along the characteristics of a trustee (e.g., ability, benevolence, and integrity), and three items to measure benevolence based on a Likert-type scale are, for example: (a) "I believe that the trustee would act in my best interest," (b) "If I required help, the trustee would do his or her best to help me," and (c) "The trustee is interested in my well-being, not just his or her own" (McKnight, Choudhury, & Kacmar, 2002, p. 355). If trust is conceptualized as an *attitude* or *behavioral intention*, other items, though similar to the three mentioned, are used to measure trust (McKnight et al., 2002).

Most behavioral studies measure trust by means of survey instruments. Thus, many studies do not measure actual trust behavior. Rather, antecedents of trust behavior (i.e., beliefs in the trustworthiness of a trustee, attitudes toward the trustee, and behavioral intentions) are measured. However, if trust is conceptualized as a behavior, it could be more advantageous to measure *actual behavior* (e.g., cooperative behavior) in a trust situation (e.g., in an economic game) rather than by asking people about their beliefs, attitudes, or intentions in a hypothetical trust setting (Fehr, 2009a).

Figure 1 (the lower part) shows that survey instruments have been used to measure trust beliefs, attitudes, and intentions, whereas the so-called trust game, an economic game, was developed to measure actual trust behavior. Moreover, it is illustrated in Figure 1 that all components of a trust situation (from perceptions of stimuli such as a trustee's facial expression to actual behavior) are based on biological factors, namely genes, hormones, and the brain (Cesarini et al., 2008; King-Casas et al., 2005; Kosfeld et al., 2005).

The *trust game* was developed to measure both trust and trustworthiness as actual behavior of players in an economic exchange game (Berg, Dickhaut, & McCabe, 1995; Camerer & Weigelt, 1988; McCabe, Rigdon, & Smith, 2003; McCabe & Smith, 2000).[1] In the mainstream version of the trust game, decision maker 1 (the trustor, DM 1) has an initial endowment of *x* monetary units (e.g., \$10, see Figure 3). First, DM 1 decides whether to share his or her endowment (e.g., \$5 for each player, then the game ends, see left path in Figure 3) or to send (a part of) it to DM 2 (the trustee) (right path in Figure 3). DM 2 observes DM 1's action and, if money was sent, decides whether to keep the amount or share (some of) it with DM 1. The experimenter triples DM 1's transfer, so that both players are better off collectively if DM 1 transfers money and DM 2 sends back a sufficient amount. Hence, in the example illustrated in Figure 3, DM 2 has two possibilities: Either to share the money (i.e., each player gets \$15) or to keep all the money (i.e., DM 1 gets \$0, DM 2 gets \$30).

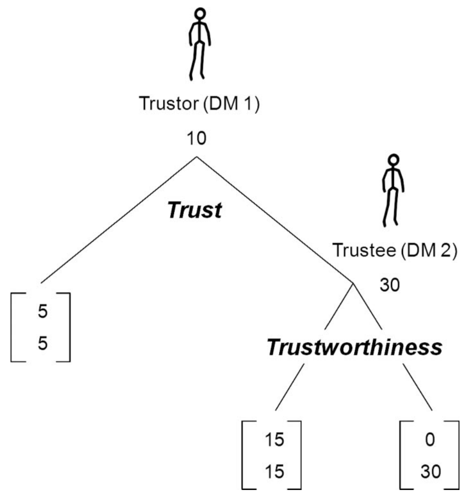


Figure 3. Structure of the trust game. *Notes:* The upper value in the square brackets indicates the trustor's payoff (decision maker 1, DM 1), the lower value the trustee's payoff (DM 2). The amounts of money and the payoff matrix in the figure are used as an example.

This game mimics a sequential economic exchange in the absence of contract enforcement institutions (Fehr & Camerer, 2007). If the game is played on a one-shot basis, DM 2 has a strong incentive to keep all the money and repay none to DM 1. If DM 1 anticipates this behavior, however, there is little reason to transfer. Consequently, if DM 1 transferred no money, then a chance for higher mutual gain would be lost. In the trust game, the amount sent by DM 1 is used as a behavioral measure for trust, and DM 2's transfer back is used as a behavioral measure for trustworthiness (see Figure 3).

[1] For a review of economic games used in neurobiological research, see Krueger, Grafman, and McCabe (2008).

The trust game has been used in many studies on the genetic, hormonal, and neurological foundations of human trust behavior. We review these investigations in the following, complemented by studies using other methods (e.g., surveys). Methodological discussions about the trust game can be found, for example, in Kugler, Connolly, and Kausel (2009) and Vilares, Dam, and Kording (2011).

2. The Biology of Trust

Genetic Level of Analysis

Survey research (Dohmen, Falk, Huffman, & Sunde, 2006) indicates that the willingness to trust other people is similar between parents and their children. This raises the question whether genetic or environmental influences, or both in conjunction, are responsible for this result. Genetic and environmental influences on behavior can be investigated either through quantitative or molecular genetic studies. Quantitative genetic methods, in contrast to molecular genetic methods, provide estimates of the relative magnitudes of omnibus genetic and environmental influences without the necessity of specifying the actual DNA sequences or environmental circumstances that provide those influences (Johnson, 2007).

To investigate whether humans are endowed with genetic variation that could account for individual differences in trust behavior, one study (Cesarini et al., 2008) applied a quantitative method based on the trust game (see Figure 3). Because monozygotic (MZ) twins share the same genes, whereas the genes of dizygotic (DZ) twins are only imperfectly correlated, MZ twins should exhibit a higher correlation in their behavior than DZ twins if genetic differences help explain the variance of trust behavior. In fact, the study of Cesarini et al. found a heritability estimate of (a) trust between 10% (Swedish subjects) and 20% (U.S. subjects) and (b) trustworthiness between 17% (U.S. subjects) and 18% (Swedish subjects). Considering these findings, Cesarini et al. (2008) concluded: "These results show that we consistently found a significant proportion of variance in trust is due to heritability" (p. 3723).

Another recent twin study (Sturgis et al., 2010) also investigated the genetic and environmental basis of trust. In contrast to the Cesarini et al. (2008) study, trust was not measured as actual behavior (trust game). Rather, a questionnaire consisting of four items was used to capture the participants' trust beliefs, namely (a) "I believe that most people are basically well-intentioned," (b) "I believe that most people will take advantage of you if you let them," (c) "I think that most of the people I deal with are honest and trustworthy," and (d) "My first reaction is to trust people." The results show that "the majority of the variance in a multiitem trust scale is accounted for by an additive genetic factor . . . the environmental influences experienced in common by sibling pairs have no discernable effect . . . [these] findings problematise the widely held view that the development of social trust occurs through a process of familial socialization at an early stage of the life course" (p. 205).

Scientific evidence demonstrates that nasally administered oxytocin (OXT, a hormone that also acts as a neurotransmitter in the brain) increases trust in humans (Kosfeld et al., 2005), highlighting the importance of this hormone for cooperative behavior (details on trust hormones are reported in the next section). Genetic research on OXT focuses on the oxytocin receptor (OXTR) gene, which (a) is located on chromosome 3p25 (Inoue et al., 1994) and (b) shows multiple variations (so-called single nucleotide polymorphisms, SNPs). These SNPs are related to differences in human perception and behavior (e.g., empathy and stress reactivity, Rodrigues, Saslow, Garcia, John, & Keltner, 2007, or autism, Wermter et al., 2010). Therefore, they have been hypothesized to play an important role in explaining individual differences in trust behavior (Reuter et al., 2009). To test this hypothesis, Reuter et al. conducted a trust game experiment with participants whose OXTR gene was screened. The results show that participants who have a particular variant of the OXTR gene exhibit more trust taking than those who exhibit the alternative variant of the gene. Reuter et al. (2009) concluded: "Our results indicate that individual differences in the proclivity to trust are influenced by variations in the OXTR gene" (p. 21).

Another study (Israel et al., 2009) also demonstrated that genetic polymorphisms for the OXTR gene are associated with human prosocial decision making behavior. In line with this finding, an investigation by Tost et al. (2010), using a multimodal imaging intermediate phenotype approach, demonstrated that a genetic variant in the OXTR gene linked to social functions predicts individual differences in brain structure and functioning, as well as personality (e.g., reward dependency, a construct associated with trust; Delgado, Frank, & Phelps, 2005).

It is important to note that a recent study found *no* significant association between nine SNPs of the OXTR gene and behavior in the trust game (Apicella et al., 2010). However, Apicella et al. write that "[g]iven that our research design only allows us to statistically reject moderate to large effect sizes, the results reported here are not inconsistent with the results of hormonal association studies involving OXT in trust and generosity and do not necessarily rule out a role for OXTR polymorphisms in explaining phenotypic variation" (p. 7).

In addition to research on the OXTR gene, there is evidence that variants in the length of a promoter region in the arginine vasopressin 1a receptor (AVPR1a) gene predict altruistic behavior. Because altruism is defined as a selfless concern for the welfare of other individuals, and hence is strongly related to one component of trustworthiness, namely benevolence (Mayer et al., 1995), this research provides additional evidence that human trust behavior is to some extent genetically predetermined. The findings of a recent review (Skuse & Gallagher, 2011) on the genetic influences on social cognition, a concept closely related to trust (Fehr, 2009b; Krueger et al., 2007), support this notion. In particular, this review identifies genetic variation in the receptors associated with OXT and AVP as a major determinant of differences in human trust behavior.

Hormonal Level of Analysis

Gene expression is not only essential for the development of brain structures and hormones (Harris, Vernon, & Boomsma, 1998; Rushton & Ankney, 2007), but is itself influenced by hormones (Harlan, 1988). Hormone regulation occurs in the hypothalamus, which in turn affects activation in the pituitary gland (Harris, 1948). These findings support our unifying framework on the biology of trust which describes genes, hormones, and brain processes as interrelated factors (see Figure 2).

In the following, we review the literature that has investigated the effects of various hormones and neurotransmitters on human trust behavior. Specifically, we discuss oxytocin, arginine vasopressin, cortisol, dopamine, testosterone, estrogen, and serotonin.

Oxytocin. Oxytocin (OXT), a neuropeptide, has been shown to influence human social behavior (Heinrichs & Domes, 2008; Heinrichs, von Dawans, & Domes, 2009). In particular, OXT plays an important role in *pro*social behaviors such as parturition, lactation, maternal attachment, and pair bonding (Donaldson & Young, 2008). Hence, OXT is associated with approach and human trust behavior.

Anatomically, OXT is synthesized in magnocellular neurons of the paraventricular and supraoptic nuclei of the hypothalamus. Moreover, it is processed from its precursor form, together with the carrier protein, along the axonal projection to the posterior pituitary, from which the peptide is secreted into the systemic circulation. Also, OXT is distributed throughout the central nervous system from smaller parvocellular neurons, influencing many neurobehavioral functions (Heinrichs, Baumgartner, Kirschbaum, & Ehlert, 2003).

Based on empirical evidence which shows that OXT plays a prominent role in facilitating various social behaviors of animals (Insel & Young, 2001), Zak and colleagues formulated the hypothesis that OXT also affects human trust behavior. Using the trust game, Zak, Kurzban, and Matzneret (2004, 2005b) found two results: First, OXT levels are, on average, 41% higher in those subjects (DM 2 in Figure 3) who receive a monetary transfer that reflects an intention of trust relative to a random monetary transfer of the same amount. Second, subjects who receive an intentional trust signal return, on average, 53% of the amount they received to the gaming partner, whereas in the random monetary condition the mean amount returned is only 18%. Thus, the perception of a signal of trust increases OXT levels, which in turn cause trustworthy behavior (i.e., the reciprocation of trust). In other words, the studies by Zak and colleagues show that when people are trusted, their brains release OXT, which in turn predicts trustworthiness. Thus, trusting behaviors depend significantly on the endogenous release of OXT.

These results suggest that trust put in another individual is likely to pay off, namely via reciprocity. But why do humans reciprocate trust even when they interact with strangers who they will never meet again? Recent explanations to understand this phenomenon are based on the concept of evolutionary psychology. This theorizing (Lieberman & Eisenberger, 2009) is based on the observation that "our emotional responses to [complex social] events rely on much of the same neural circuitry that underlies the simplest physical pains and pleasures" (p. 890) and "the brain may treat abstract social experiences [such as trust] and concrete physical experiences as more similar than is generally assumed" (p. 891).

This observation can be explained by two evolutionary phenomena (Lieberman & Eisenberger, 2009): First, mammalian newborns depend on other humans for survival, because they are relatively immature at birth. Hence, the survival of a newborn, and thus the survival of mankind in general, depends on the social bond between a newborn and other humans, which implies trust. Second, the division of basic human activities among group members (e.g., one individual takes care of food acquisition, another one of protection from adversarial groups, and another one of care for offspring) has turned out to guarantee survival, while a failure to divide the various human activities did not. This division of activities among group members, however, is associated with cooperative behavior, which also implies trust. Therefore, trust contributes to survival from an evolutionary perspective. In line with this argumentation, a recent study by De Dreu et al. (2010) found that intranasal administration of OXT modulates trust toward members of one's own group, but not toward members of competing out-groups.

In another trust game experiment, Kosfeld et al. (2005) administered—via nasal spray—either OXT or a placebo to subjects (DM 1 in Figure 3). As the results show, OXT considerably increases the trustors' trust into the trustees. Out of the 29 subjects, 13 (45%) in the OXT group showed the maximal trust level (i.e., they sent all the money), whereas only six of the 29 subjects (21%) in the placebo group showed maximal trust. Thus, the Kosfeld et al. experiment shows that exogenously administered OXT increases trust.

This result, however, allows for a number of causal interpretations (e.g., Fehr, Fischbacher, & Kosfeld, 2005). First, did OXT increase the trustors' trust by creating more optimistic beliefs about the trustworthiness of the trustees? Second, did OXT make trustors more generous? Third, did it make trustors more risk/ambiguity-seeking? Or fourth, did OXT influence betrayal aversion? Because the experiment of Kosfeld et al. (2005) controlled for these alternative explanations, it was possible to find out that OXT helps to overcome betrayal aversion (Bohnet, Greig, Herrmann, & Zeckhauser, 2008; Damasio, 2005).[2] Thus, it can be theorized that OXT affects the neurobiological mechanisms that underlie social preferences (Fehr, 2009a, 2009b). However, although the effect of OXT on generosity was ruled out in the Kosfeld et al. (2005) study, other investigations (Barraza & Zak, 2009; Zak, Stanton, & Ahmadi, 2007) identified a positive correlation between OXT and generosity. Considering this, current research suggests that OXT affects both betrayal aversion and generosity.

When OXT levels are increased, people become more trusting (e.g., Kosfeld et al., 2005). Mikolajczak et al. (2010), however, raised the following question: Does OXT increase people's trust in anybody, or can contextual cues of unreliability override the effects of OXT? The results of their experiment show that people given OXT rather than placebo trust others more to the extent that trustees were displayed in a neutral or positive fashion. However, when the trustees were displayed in a negative fashion, OXT did not influence trust. This result suggests that the effect of OXT on trust may be moderated by the perceived risk inherent to the interaction (Mikolajczak et al., 2010).

Another study (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007) investigated the effect of intranasally administered OXT on mind-reading ability. Mind-reading is the ability to infer the internal states of other actors (e.g., intentions, thoughts, and feelings) to predict their behavior (also known as theory of mind, TOM), and the underlying inference process is commonly referred to as mentalizing (Frith & Frith, 2003; Premack & Woodruff, 1978; Singer, 2009). The study by Domes et al. shows

[2] Aimone and Houser (2011) show that groups including betrayal-averse agents may achieve higher levels of reciprocity and more profitable social exchange than groups lacking betrayal aversion. These findings provide evidence on the benefits of betrayal aversion in trust situations.

that OXT improves the ability to infer the mental state of others from social cues of the eye region.

In general, mentalizing is a fundamental cognitive process in trust situations because the decision to trust involves thinking about an interaction partner's intentions, thoughts, and feelings to infer his or her trustworthiness (Fehr, 2009b; Krueger et al., 2007). Thus, in addition to its effects on betrayal aversion and generosity, OXT also has an effect on mentalizing. This further substantiates the prominent role of OXT for the establishment of interpersonal trust.

Arginine vasopressin. Arginine vasopressin (AVP), another neuropeptide, has also been shown to influence human social behavior (Heinrichs & Domes, 2008; Heinrichs et al., 2009). AVP is associated with a variety of male-typical social behaviors including aggression, territoriality, and enhanced stress-responsiveness (Donaldson & Young, 2008; Heinrichs & Domes, 2008). Consequently, AVP is hypothesized to be associated with avoidance and distrust behavior rather than with approach and trust behavior.

It is of importance for the neurobiological foundations of trust that research has already investigated the effects of nasally administered AVP on human facial responses related to social communication. One study (Thompson, Gupta, Miller, Mills, & Orr, 2004), for example, was designed to determine if AVP administration would influence cognitive, autonomic, as well as somatic responses to social stimuli important for agonistic communication in humans (agonistic is used as a synonym for aggressive in this context). Specifically, the study tested the effects of intranasal AVP administration on attention toward emotionally expressive facial expressions, as well as on heart rate, skin conductance, and corrugator supercilii electromyograms (corrugator EMG) in response to these social stimuli. The results show that AVP did not affect attention toward emotionally neutral, happy, and angry facial expressions. Also, AVP did not influence autonomic arousal in response to the three facial expressions. However, AVP did selectively enhance the corrugator EMG responses evoked by emotionally neutral facial expressions, making them similar in magnitude to responses evoked by angry facial expressions in control subjects.

In their discussion of the results, Thompson et al. (2004) argue that "[b]ecause this muscle group is involved in agonistic communication, these results suggest that AVP may influence aggression in human males by biasing individuals to respond to emotionally ambiguous social stimuli as if they were threatening/aggressive" (p. 35). Considering this and similar research findings (e.g., it was found that AVP decreases perceptions of the friendliness of faces, Thompson, George, Walton, Orr, & Benson, 2006), AVP can be considered as the adversary of OXT. Thus, AVP is associated with distrust, while OXT is associated with trust. In line with this reasoning, research has also shown that AVP is related to increased vigilance and anxiety (Carter, 2007; Murgatroyd et al., 2004).

Cortisol. Glucocorticoids, and cortisol (COR) in particular, are secreted by the adrenal cortex. Hence, they are also referred to as corticosteroids. These corticosteroids have been shown to have a number of neural and behavioral effects which are important in trust situations, such as learning and memory. Research has found, for example, that stress-induced COR elevation impairs social memory (Takahashi et al., 2004), in particular person recognition memory based on facial information (Rimmele et al., 2009), which is crucial for trust decisions in face-to-face settings (Winston et al., 2002). Moreover, psychosocial constructs such as optimism, self-mastery, self-esteem, and extraversion are associated with low activation in fear-related brain areas, which in turn results in attenuated COR responses to stress (Taylor et al., 2008).

Other studies (Heinrichs et al., 2003; Takahashi et al., 2005) investigated the direct relationship between trust and COR elevation induced by a social stress test. The results revealed a significant negative correlation between the two factors. That is, subjects with higher degrees of trust disposition have reduced social stress-induced COR response. Because (a) the stress hormone COR is synthesized in response to activation in the amygdala (Tillfors et al., 2001), and (b) lesions of the central nucleus of the amygdala decrease stress-induced hormone release (Prewitt & Herman, 1994), amygdala activation in healthy people is associated with distrust (Adolphs, Tranel, & Damasio, 1998; Dimoka, 2010; Takahashi et al., 2005).

There is also a link between neuropeptides and corticosteroids. Intranasal administration of AVP, for example, has been shown to increase salivary COR levels in a social stress test (de Winter et al., 2003; Ebstein et al., 2009; Patchev & Almeida, 1995), while intranasal administered OXT reduced the COR response to social stress (Ditzen et al., 2009; Heinrichs et al., 2003).

Dopamine. There is evidence for a relation between hypothalamo-pituitary-adrenal (HPA) axis responses and the dopaminergic system (Rostene et al., 1995). It has been reported that selective destruction of mesencephalic dopaminergic neurons in rats leads to decreased levels of corticosterone (Casolini et al., 1993). Also, it has been shown that COR regulates dopamine (DOP; Barrot et al., 2000).

DOP is the main neurotransmitter associated with reward processing (Ikemoto & Panksepp, 1999; Schultz, 2000, 2002, 2006; Schultz, Dayan, & Montague, 1997). Because the goal of trusting another individual is to realize a reward (e.g., through a beneficial cooperation), it is no surprise that OXT has been shown to enhance DOP levels (Shahrokh, Zhang, Diorio, Gratton, & Meaney, 2010; Zak, 2011). Altogether, the important role of reward related neurobiological mechanisms in trust situations is stressed in multiple studies (e.g., King-Casas et al., 2005; Krueger et al., 2007)—Zak et al. (2005b) write in this context: "[T]he findings here should be considered in light of the fMRI study of Rilling et al. (2002), who show significant activity in ventromedial regions rich in dopamine receptors during cooperative behaviors. In the prairie vole, the nucleus accumbens is dense in OT [oxytocin] receptors . . . and OT appears to be critical for linking social signals to ventromedial reward circuits" (p. 526). In line with this argumentation, a recent review (Baskerville & Douglas, 2010) indicates the existence of a positive interaction between DOP and OXT.

Testosterone. Sex hormones also play a key role in human social behavior. Perception of distrust is correlated with an increase of dihydrotestosterone, a metabolite of the male sex hormone testosterone (TES) (Zak, Borja, Matzner, & Kurzban, 2005a). Moreover, elevated TES also causes people to behave antisocially (Zak et al., 2009). Based on the knowledge that TES is associated with competition and dominance, and therefore considered as an inhibitor of sociality, one recent study (Bos, Terburg, & Van Honk, 2010) tested whether TES has antagonistic characteristics with OXT. The results of the study show that compared with the placebo, TES significantly decreased interpersonal trust. Hence, the results confirm the predicted antagonistic characteristics with OXT. Moreover, additional analyses of the data revealed that this effect is determined by those who give trust easily, so-called naïve humans. Bos et al. (2010) conclude that "testosterone adaptively increases social vigilance in these trusting individuals [in particular the naïve humans] to better prepare them for competition over status and valued resources" (p. 1). Further support for the theorizing of Bos et al. (2010) comes from Johnson and Breedlove (2010) who argue that TES "may reduce interpersonal trust by acting on vasopressinergic neurons in the amygdala to increase communication to brainstem systems that activate fearful responses" (p. 11149).

An animal study (Arsenijevic & Tribollet, 1998), however, demonstrated that TES treatment of aging rats can restore normal adult levels of OXT receptor binding in specific brain areas. Thus, this study suggests a positive correlation between TES and OXT, thereby challenging the results of other research findings.

In another study (Zethraeus et al., 2009) that investigated the effects of sex hormones (e.g., TES) on various forms of social behaviors (e.g., trust, altruism, fairness), subjects were allocated to four weeks of treatment with sex hormones or a placebo. At the end of the treatment period, the subjects participated in experiments that were designed to measure the various forms of social behaviors. In essence, no significant effect of TES on any of the social behaviors was found. This result and findings of similar studies (e.g., Eisenegger, Naef, Snozzi, Heinrichs, & Fehr, 2010; Sanchez-Pages and Turiegano, 2010) substantiate the notion that the relation between TES and OXT is still not well understood. However, because studies with human subjects, in contrast to animal studies, found a negative relation between TES and OXT, we lend more weight to this theoretical perspective. Nevertheless, future studies should address the relation between TES and OXT to gain further insights.

Estrogen. With respect to the female sex hormone estrogen (EST), it has been reported that EST facilitates OXT uptake by facilitating receptor binding and increasing the number of OXT receptors (Verbalis, 1999). In line with this finding, one study (Zak & Fakhar, 2006) analyzing data on biological, social, and environmental factors associated with hormone levels for a sample of 41 countries found that particular environmental conditions in some nations may be conducive to higher trust levels. In particular, this study found that nations whose citizens consume more food containing phytoestrogens (i.e., a diverse group of naturally occurring nonsteroidal plant compounds which may have similar effects as endogenous EST) exhibit higher levels of trust disposition. Thus, EST and OXT are positively correlated (Zak & Fakhar, 2006).

Serotonin. Serotonin (SER), a neurotransmitter mainly produced in the raphe nuclei, has multiple behavioral functions, in particular those related to social status, sexual behavior, as well as aggression and mood regulation. There is evidence that SER deficiency is associated with violent and aggressive behavior, while high levels of AVP facilitate aggressive behavior (Ferris et al., 1997). Thus, SER antagonizes AVP activity in the brain (Ferris & Delville, 1994).

SER depletion has been found to result in a higher rejection rate of unfair offers in an economic game (Crockett, Clark, Tabibnia, Lieberman, & Robbins, 2008), thereby indicating the role of SER for the control of negative emotions which also play an important role in trust situations (Winston et al., 2002). In line with this result, one study (Wood, Rilling, Sanfey, Bhagwager, & Rogers, 2006) found that subjects deprived of L-tryptophan, and consequently lower SER levels in the brain, are less cooperative in a repeated economic game than control subjects. In line with this finding, a recent article by Zak (2011) indicates that OXT release causes synaptic SER to rise, which in turn leads to calmness and positive mood by binding to 5-HT1A (SER) receptors in the temporal and prefrontal cortices.

Depression is also known to be the result of a serotonin deficiency. Also, depressive patients have been shown to have higher serum COR levels (Cowen, 2002). This suggests a link between SER and COR. Stress is known to activate the serotonergic system (McKittrick, Blanchard, Hardy, & Blanchard, 2009). Hence, considering the prominent role of COR for stress perceptions, a negative correlation between SER and COR with a positive feedback loop seems to exist. In line with this finding, research found that touch is associated with increased levels of SER (Field, Grizzle, Scafidi, & Schanberg, 1996), which have been demonstrated to reduce stress reactivity in humans (Hanley & Van de Kar, 2003). Finally, one investigation (Skuse & Gallagher, 2011) indicates that SER desensitizes the AVP receptor, which could increase avoidance and distrust behavior.

Summary of hormonal level of analysis. Figure 4 summarizes the results of our hormone review. The following list reflects the most important findings:

- Estrogen (EST), oxytocin (OXT), dompamine (DOP), and serotonin (SER) are typically associated with approach behavior and trust, whereas arginine vasopressin (AVP), cortisol (COR), and testosterone (TES) are more strongly associated with avoidance behavior and distrust.
- OXT, in contrast to the other hormones and neurotransmitters, seems to play a more central role for human trust behavior, as this substance has connections to five other substances, whereas the other substances have fewer connections.
- OXT is mainly influenced by sex hormones (i.e., EST and TES).
- The hormones and neurotransmitters make possible inferences regarding the psychological processes associated with a trust decision. In particular, reward processing (DOP), fear/uncertainty/stress processing and memory (COR), as well as mentalizing and processing of social preferences such as betrayal aversion and generosity (OXT) are important psychological processes involved in interpersonal trust situations.

Brain Level of Analysis

Genes influence the anatomy of the brain and its processing mechanisms (Rushton & Ankney, 2007; Rushton & Jensen, 2008). Moreover, the release of hormones is controlled by the brain, and, in turn, hormones influence the activation of particular brain regions via corresponding receptors (Loup et al., 1991). Consequently, a model of the biological determinants of human trust behavior has to consider the brain with its multifarious structures. In the following, we

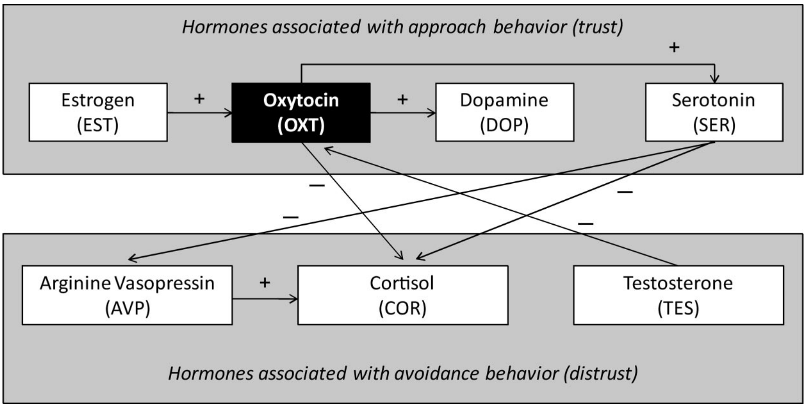


Figure 4. Relations among important trust-relevant hormones and neurotransmitters. *Notes:* The black box indicates that oxytocin (OXT) is the hub in the system of trust- and distrust-related hormones.

present empirical evidence showing that human trust behavior is associated with activation in specific brain areas.

A detailed look at the fMRI literature reveals that three different experimental paradigms have been used so far to study the neural correlates of trust. First, several studies have used the *trust game* as a paradigm. In this article, we discuss the following trust game studies: Baumgartner, Heinrichs, Vonlanthen, Fischbacher, and Fehr (2008); King-Casas et al. (2005); Delgado et al. (2005), and Krueger et al. (2007). Second, one study (Winston et al., 2002) used *human faces* with a varying degree of trustworthiness as stimuli. Third, two studies (Dimoka, 2010; Riedl, Hubert, & Kenning, 2010) used *eBay Web sites* as stimulus material in order to take into account the increasing trend toward computer-mediated interactions among humans. We structure the following section along these three experimental paradigms.

Trust game paradigm. One trust game experiment (Baumgartner et al., 2008) combined intranasal administration of OXT and fMRI. The results of the study revealed that subjects who had received OXT showed no change in their trust behavior after they learned that their trust had been breached several times, while subjects receiving a placebo decreased their trust. This result is in line with the findings of other hormone studies (Kosfeld et al., 2005; Zak et al., 2004, 2005b). Moreover, as the brain scans show, this difference in trust adaptation is associated with a specific reduction in activation in the amygdala, the midbrain regions, and the dorsal striatum in subjects receiving OXT, suggesting that neural systems mediating fear processing (amygdala and midbrain regions) and behavioral adaptations to feedback information (caudate nucleus, which is a part of the striatum) modulate the effect of OXT on trust (Baumgartner et al., 2008; Kirsch et al., 2005).

In addition to their main results, Baumgartner et al. (2008) list a number of additional brain regions activated in the trust game, namely putamen (a part of the striatum), thalamus, insular cortex, anterior cingulate cortex (ACC), inferior temporal gyrus, precuneus/posterior cingulum, superior parietal gyrus, postcentral gyrus, and the red nucleus. At least four regions on this list are hypothesized to play a crucial role in the neural implementation of mental processes that are important in trust situations.

First, activation in the striatum and thalamus is associated with reward processing and reward anticipation (Komura et al., 2001; O'Doherty et al., 2004; Schultz, 2006; Tricomi, Delgado, & Fiez, 2004). Second, activation in the insular cortex is related to perception of faces that express negative emotions (Phillips et al., 1997, 1998). Moreover, insular activation is associated with perception and processing of uncertainty, risk, and ambiguity (Critchley, Mathias, & Dolan, 2001; Huettel, Stowe, Gordon, Warner, & Platt, 2006; Kuhnen & Knutson, 2005; Krain, Wilson, Arbuckle, Castellanos, & Milham, 2006). Third, activation in the ACC is associated with cognitive conflict (Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Botvinick, Cohen, & Carter, 2004). Considering these research findings, it can be hypothesized that striatum and thalamus activation plays a significant role in trust perceptions, whereas activation in the insular cortex is crucial for distrust perceptions. ACC activation, in contrast, is not trust or distrust specific. Rather, it is related to the comparison of the potential benefits and risks associated with a trust decision, thereby leading to cognitive conflict.

Using a multiround version of the trust game, King-Casas et al. (2005) found that reciprocity expressed by one player strongly predicts future trust expressed by the other player. Examining the trustees' brain scans, the study found that activity in the caudate nucleus was greatest when the trustor showed benevolent behavior, and most subdued when the trustor showed malevolent behavior. Moreover, activity in the caudate rose and fell with changes in the amount of money trustees returned to their trustors on the subsequent round, suggesting that this brain circuit receives or computes information about both the fairness of a social partner's decision and the intention to repay that decision with trust. Thus, activation in the caudate nucleus signals trust and/or learning of a partner's trustworthiness (King-Casas et al., 2005; Miller, 2005). In addition to caudate activity, the study revealed further brain regions activated in the trust game, namely the thalamus, ACC, inferior frontal sulcus, superior frontal sulcus, inferior/superior colliculi, and middle cingulate cortex.

In another fMRI study (Delgado et al., 2005), participants played a trust game with three hypothetical partners depicted as having moral character that was good, bad, or neutral. As the results show, participants chose to be more cooperative with the morally good partner. Moreover, caudate nucleus activation differentiated between positive and negative feedback, but only for the neutral partner. Notably, it did not do so for the good and bad partners, suggesting that prior social and moral perceptions can diminish reliance on feedback mechanisms in the caudate nucleus. Overall, the study provides evidence that moral and social perceptions can modulate neural mechanisms associated with feedback and reward processing and cognitive control in trust situations (Delgado et al., 2005; Singer, Kiebel, Winston, Dolan, & Frith, 2004).

In addition to the caudate, Delgado et al. (2005) found that the ventral portions of the striatum also play an important role in trust situations, because they are crucial for making predictions and anticipating the outcome of risky decisions. Finally, the study of Delgado et al. (2005) revealed high activation of the insular cortex and the cingulate cortex in trust situations—the former brain structure is important for the perception and processing of negative emotions (e.g., uncertainty), whereas the latter is important for the processing of cognitive conflict. This result is in line with the findings reported in the Baumgartner et al. (2008) study.

In a hyperfunctional MRI study (Krueger et al., 2007), two strangers interacted with one another in a sequential reciprocal trust game, while their brains were simultaneously scanned. The findings of the study suggest that the paracingulate cortex is critically involved in building a trust relationship by inferring another person's intentions to predict subsequent behavior (Bhatt & Camerer, 2005). In addition to the paracingulate cortex—which was identified as a mentalizing brain area not only by Krueger et al. (2007), but also by Gallagher, Jack, Roepstorff, and Frith (2002) and McCabe, Houser, Ryan, Smith, and Trouard (2001), the medial prefrontal cortex has also been found to be involved in thinking about other people's mental states (McCabe et al., 2001; Siegal & Varley, 2002; Stuss, Gallup, & Alexander, 2001). In particular, it is activated when making decisions and choices based on calculative expectations of what others will do (Bhatt & Camerer, 2005). Moreover, a near-infrared spectroscopy (NIRS) study (Yanagisawa et al., 2011) found that activity in the right ventrolateral prefrontal cortex (rVLPFC) is positively correlated with general trust. Because the level of general trust was negatively correlated with self-reported social pain during a social exclusion situation in this experiment, Yanagisawa et al. conclude that "rVLPFC activity mediated the relationship between general trust levels and social pain" (p. 190). In line with these findings, a recent review by

Rilling and Sanfey (2011) highlights the role of the prefrontal cortex in social decision-making.

Finally, a further brain region associated with mentalizing is the posterior superior temporal sulcus (Frith & Frith, 2003). This brain region is of particular importance in perceiving and processing biological motion of other species (Frith & Frith, 2010). Thus, in real-life situations (in contrast to static contexts like the trust game), this brain region is also likely to be activated in interpersonal trust situations.

The study of Krueger et al. (2007) revealed further brain regions associated with trust behavior: the ventral tegmental area, a region linked to the evaluation of expected and realized reward (Schultz et al., 1997), and the septal area, together with the adjoining hypothalamus, a limbic region that has been demonstrated to modulate various aspects of social behavior including social memory and learning (Numan, 2000). Because trust situations do not only exist on the basis of one-shot interpersonal interactions, but also on the basis of repeated interactions, memory and learning are crucial mental processes involved in human trust (King-Casas et al., 2005). Also, it has been shown that the septal area plays a crucial role in the release of several trust-related neuropeptides such as OXT. In addition, this area contains receptors for OXT and other neuropeptides (Loup et al., 1991; Powell & Rorie, 1967), and it is strongly connected with the hippocampus, which also has OXT receptors (Heinrichs, Meinlschmidt, Wippich, Ehlert, & Hellhammer, 2004), substantiating the notion that the septal area is an important brain region associated with human trust.

Human faces as a paradigm. An fMRI experiment conducted by Winston et al. (2002) found that the amygdala and insula were significantly more activated when subjects viewed faces that they rated as most untrustworthy, suggesting that these two brain regions associate perception of a face with an emotional response of distrust (Adolphs, 2002; Winston et al., 2002). Moreover, the study found activation in mentalizing brain areas (superior temporal sulcus). Based on these results, Winston et al. (2002) concluded that "social judgments about faces reflect a combination of brain responses that are stimulus driven, in the case of the amygdala, and driven by processes relating to inferences concerning the intentionality of others" (p. 281).

With respect to amygdala activation, three case studies (Adolphs et al., 1998, 2005) found that patients with complete bilateral amygdala damage judged other people to look more trustworthy and more approachable than did normal viewers or patients with brain damage in other areas, providing additional empirical evidence that the amygdala is associated with distrust. Other empirical studies found that this region is also activated by viewing faces that show certain emotional expressions, notably fear (Morris et al., 1996), and that it is activated when subjects view faces of people from another race (Hart et al., 2000; Phelps et al., 2000). In line with these findings, a recent lesion study (Koscik & Tranel, 2011) found that participants with unilateral damage to the amygdala displayed increased benevolent behavior in a multiround trust game, and specifically, they increased trust in response to betrayals. Thus, research indicates unambiguously that the amygdala is necessary for expressing normal interpersonal trust.

In a comment on the Winston et al. (2002) article, Adolphs (2002) explains that a detailed look at the stimulus material and the experimental designs reveals that "expressions of anger or sadness were negatively correlated with trustworthiness ratings, and happiness was positively correlated" (p. 193). Thus, the trustworthiness assigned to a particular face seems to depend on the emotion conveyed by the face (Todorov, 2008; Todorov, Baron, & Oosterhof, 2008).

However, such a conclusion cannot explain why different viewers of faces assign differential trustworthiness ratings. Adolphs (2002) also offers an explanation for this. In essence, he argues that trustworthiness ratings are a function of both the features of the stimuli (e.g., the emotion conveyed by a face) and the personalities and autobiographies of the viewers. Bearing this in mind, future research should not only manipulate the features of the stimuli. Rather, future studies should also recruit groups of subjects with varying personalities. For example, because autistic people have deficits in mentalizing (Baron-Cohen, Leslie, & Frith, 1985; Frith & Frith, 2010), which in turn has been shown to be associated with OXT levels (Domes et al., 2007) and activation in specific brain regions such as the medial prefrontal cortex (Frith & Frith, 2006), a comparison of behavior and corresponding brain activation patterns and hormone levels between autistic people versus healthy controls in trust game experiments (or other experimental paradigms) could reveal new insights into the neurobiological mechanisms of the formation of trustworthiness judgments.

It is important to outline that it has not only been demonstrated in a face evaluation task (Winston et al., 2002) that automatic and unconscious information processing plays a crucial role in trust decisions. Based on their investigation in which they administered OXT to fMRI subjects, Baumgartner et al. (2008) for example, write: "[D]ifferences in brain activation between placebo and OT [oxytocin] subjects were only observed in subcortical structures as the amygdala, the midbrain, and the striatum. Those brain structures have each been associated with automatic and intuitive . . . or even unconscious processes" (p. 646). Prominent scholars such as Damasio (2005) and Zak (2003) support the importance of automatic and unconscious information processing in trust situations; the latter, for example, writes: "[T]rust is not a calculative activity . . . but a visceral sense that one has that a person can be trusted or not" (p. 21). Altogether, these findings indicate that a considerable proportion of neural processing in trust situations seems to be associated with automatic, emotional, and unconscious rather than deliberate, neutral, and conscious information processing.

Web sites as a Paradigm. So far, we have discussed the neurobiological foundations of trust based on the assumption that a trustor and a trustee interact with each other in a face-to-face setting (see Figure 1). During the past thousands of years, interaction among humans took place in such face-to-face environments. Today, however, increasingly more interactions among humans are computer-mediated. Information and communication technologies, in particular the Internet, have dramatically changed the context in which interactions among humans take place. In today's society, approximately two billion people use the Internet for communication and interaction, and this number is increasing on a daily basis (Internet World Stats, 2010).

As a result of this shift from face-to-face interaction to Internet-based communication and transactions, traditional human-human interactions are becoming increasingly more often computer-mediated. The use of computers and the Internet, however, usually increases a perception of complexity and uncertainty in both social and economic exchange (Pavlou, Liang, & Xue, 2007). Thus, trust as a construct is even more important in Internet environments than in traditional contexts in establishing interaction and cooperation among humans. A major reason for the increasing complexity and uncertainty in Internet environments is that during an interaction, a partner's face, which is typically not observable in computer-mediated interaction, serves the interpersonal function of allowing one person to predict another's personality traits and behavior (Knutson, 1996; Oosterhof & Todorov, 2008). This, in turn, has been demonstrated to apply in particular to trustworthiness predictions (Todorov, 2008; Winston et al., 2002). Hence, prediction of another person's personality traits and behavior is more difficult in computer-mediated interaction than in traditional face-to-face interaction.

Against this background, recent neuroimaging studies have begun to investigate the neural correlates of trust in computer-mediated interactions among humans. One fMRI study (Dimoka, 2010) used Web sites of eBay feedback profiles with varying levels of trustworthiness (manipulation based on the ratio between positive, neutral, and negative feedback comments) as stimuli to trigger brain activation (feedback profiles can be used by buyers to evaluate the quality of a transaction they have conducted with a seller; hence, sellers have an incentive to act in a trustworthy manner to build up a good reputation). In essence, the study found that trust is associated with brain areas linked to anticipating rewards (caudate nucleus), predicting the behavior of others (anterior paracingulate cortex), and calculating uncertainty (orbitofrontal cortex). Distrust, in contrast to trust, was found to be associated with brain areas linked to intense negative emotions (amygdala) and fear of loss (insular cortex).

Another brain imaging study (Riedl et al., 2010) used Internet offers of eBay sellers as stimulus material. Specifically, Toulmin's (1958) model of argumentation was applied to develop product description texts with varying degrees of trustworthiness. The results of the study show that the processing of trustworthy eBay offers activates reward processing areas (striatum and thalamus) and mentalizing areas (prefrontal regions and cingulate cortex). In contrast, the processing of untrustworthy eBay offers activated regions associated with perception of uncertainty, in particular the insular cortex.

Summary of brain level of analysis. Taken together, the fMRI literature on human trust behavior has identified a number of crucial brain regions. We summarize these brain regions in Table 1. A marked cell in the table indicates that a study has found this brain region activated in a trust-related task. Table 1 shows that the neural correlates of trust are not so different at all across the three different experimental paradigms (trust game, human faces, eBay Web sites). Thus, our review provides evidence that the neural correlates of trust are relatively independent of the specific trust situation.

In Table 1, we also assign the brain regions identified in our review to five classes of mental processes that are crucial in trust situations:

- *Reward Processing* is associated with activation in the striatum and thalamus.
- *Uncertainty, risk, ambiguity, and fear processing* is associated with activation in the amygdala, insular cortex, as well as the hippocampus and parahippocampus gyrus.
- *Memory* is associated with activation in the amygdala, as well as hippocampus and parahippocampus gyrus.
- *Processing of cognitive conflict* is associated with activation in the cingulate cortex (in particular in the ACC).
- *Mentalizing and deliberate thinking* are two cognitive processes that are associated with activation in the frontal cortex.

A look at the five classes of mental processes reveals that the first class (reward) is of positive valence. Thus, activation in the corresponding brain regions is hypothesized to result in trust, approach, and cooperative behavior. In contrast, the second class (uncertainty, risk, ambiguity, fear) is of negative valence. Thus, activation in the corresponding brain regions is hypothesized to result in distrust, avoidance, and withdraw behavior. The other three classes (memory, cognitive conflict, as well as mentalizing and deliberate thinking) are neither of positive nor negative valence.

3. The Biology of Trust: Toward a Unified View

In section 2, we reviewed the biologically oriented literature on trust along three levels of analysis (genes, hormones, and the brain). The major goal of this discussion was to outline the relation between trust behavior and each of the three biological factors. However, as represented by the arrows in Figure 2, the three biological factors are also interrelated. In section 2, we only touched on these interrelationships. In this section, we make the relations more explicit to derive a unified view on the biology of trust. This unified view is illustrated in Figure 5 in the form of a conceptual model.

In the following, we discuss our conceptual model of the biology of trust (see Figure 5). Every human has a certain genetic predisposition that also concerns the oxytocin receptor (OXTR) gene and the arginine vasopressin 1a receptor (AVPR1a) gene. The characteristics of these genes affect two hormones, namely OXT (Cesarini et al., 2008; Reuter et al., 2009) and AVP (Knafo et al., 2008). While high levels of OXT lead to trust behavior, high levels of AVP typically result in distrust behavior (Heinrichs & Domes, 2008; Heinrichs et al., 2009). Hence, OXT antagonizes AVP (Donaldson & Young, 2008).

As illustrated in our model, OXT is influenced by sex hormones. First, the female sex hormone EST facilitates OXT uptake (Verbalis, 1999). Thus, EST positively affects OXT. Second, the male sex hormone TES has antagonistic characteristics with OXT (Bos et al., 2010). Thus, TES negatively affects OXT. (Note that the relation between TES and OXT is currently less well understood than the relation between EST and OXT, see section Hormonal Level of Analysis.)

OXT has been shown to enhance DOP levels, which play a crucial role for reward processing (Schultz, 2000, 2002, 2006; Schultz et al., 1997). This positive relation between the two hormones is accompanied by a corresponding relation on the cognitive and behavioral levels, because the ultimate goal of trusting another

Table 1

FMRI Studies On Human Trust Behavior and Associated Brain Regions

	Baumgartner et al. (2008)	Delgado et al. (2005)	King-Casas et al. (2005)	Krueger et al. (2007)	Winston et al. (2002)	Dimoka (2010)	Riedl et al. (2010)
Sample size (female/male)	49 (0/49)	12 (5/7)	96 (n.a.)	44 (22/22)	14 (6/8)	15 (6/9)	20 (10/10)
Experimental paradigm (stimulus)	Trust game	Trust game	Trust game	Trust game	Faces	eBay websites	eBay websites
Mental processes/brain regions							
Reward							
Striatum	●	●	●			●	●
Thalamus	●		●		●		●
Uncertainty, risk, ambiguity, fear/Memory							
Amygdala*	●				●	●	
Insular cortex	●	●			●	●	●
Hippocampus, parahippocampus gyrus*		●			●		●
Cognitive conflict							
Cingulate cortex	●	●	●	●		●	●
Mentalizing and deliberate thinking							
Frontal cortex				●		●	●

Note. The table lists brain regions that are mentioned in at least three of the seven studies. The brain regions are categorized along classes of mental processes that are important in trust situations: (a) reward processing, (b) uncertainty, risk, ambiguity, and fear processing, (c) processing of cognitive conflict, and (d) mentalizing and deliberate thinking. Note that we do not claim that a particular brain region is exclusively associated with one mental process. Rather, the relation between brain regions and mental processes is best described by a "many-to-many-concept." That is, activation in a specific brain region might be associated with several mental processes, and a specific mental process might be associated with activation in several brain regions (Price and Friston 2005). The asterisk (*) indicates that the amygdala, together with the hippocampus and parahippocampus gyrus, is also a crucial brain structure for emotion-based memory and learning, which are of particular importance for trust decisions independent of the experimental paradigm. Finally, note that although five out of the seven studies used a mixed-gender sample (with the exception of the study by Baumgartner et al., and with information unavailable in King-Casas et al.), only the Riedl et al. study has a focus on gender differences in brain activation in trust decisions.

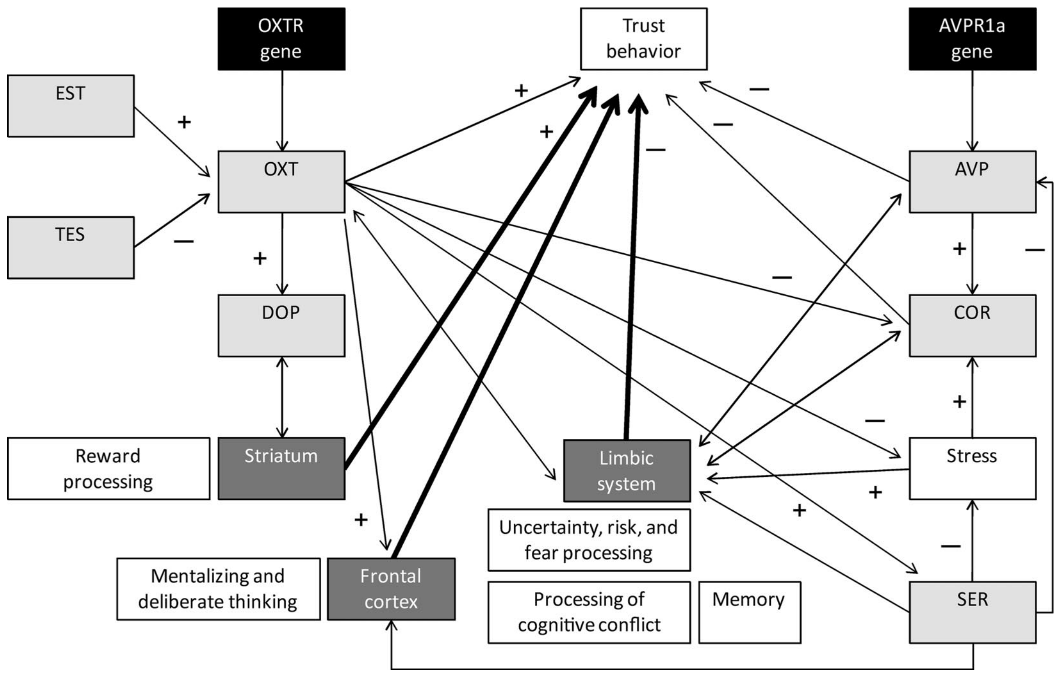


Figure 5. Conceptual Model of the Biology of Trust. *Notes:* AVP: Arginine Vasopressin, AVPR1a: Arginine Vasopressin 1a Receptor, COR: Cortisol, DOP: Dopamine, EST: Estrogen, OXT: Oxytocin, OXTR: Oxytocin Receptor, SER: Serotonin, TES: Testosterone. Black boxes indicate a genetic level of analysis. Light gray boxes indicate a hormonal level of analysis. Dark gray boxes indicate a brain level of analysis. White boxes indicate a behavioral level of analysis (trust behavior and stress), as well as cognitive constructs (reward processing; uncertainty, risk, and fear processing; processing of cognitive conflict; mentalizing and deliberate thinking). Plus (+) and minus (−) signs indicate the hypothesized valence of the relation (positive or negative). Direction of the arrows indicates the hypothesized causality. A two sided-arrow indicates interplay among two variables. The arrows in bold illustrate that the influences of genes and hormones on trust behavior are mediated by activation in specific brain areas. Thus, a direct link between a hormone and trust behavior indicates that research has investigated this relation without considering brain activity as a mediator (e.g., OXT→Trust behavior, Kosfeld et al., 2005). Anatomical connections between the brain areas (striatum, limbic system, and frontal cortex) are not illustrated to facilitate clarity of the model.

person (OXT) is the realization of a reward (DOP; Shahrokh et al., 2010). In line with this argumentation, a recent study (Theodoridou, Rowe, Penton-Voak, & Rogers, 2009) found that nasally administered OXT increases perceived facial trustworthiness and attractiveness of unfamiliar people, and this in turn may positively affect reward perceptions.

DOP is correlated with reward processing in the brain, and DOP is released by a specific part in the striatum, the nucleus accumbens (Schultz et al., 1997). Hence, an interplay between the striatum and DOP does exist. In line with our theorizing (i.e., OXT→DOP↔Striatum), a number of brain imaging experiments found activation in the striatum, and also in other reward areas like the thalamus, in trust-related tasks (see Table 1). Therefore, one major conclusion is that activation in the striatum typically leads to trust behavior (Zak, 2011).

As illustrated in Figure 5, OXT positively affects the ability to infer the internal states of other people—for example their intentions, thoughts, and feelings—to predict their behavior (Domes et al., 2007). This mentalizing process, together with deliberate thinking, is associated with activation in the frontal cortex, in particular in the medial frontal cortex (Amodio & Frith, 2006; Stuss et al., 2001). Importantly, mentalizing is a fundamental cognitive process in trust situations because the decision to trust implies thinking about an interaction partner's intentions to infer his or her trustworthiness (Fehr, 2009b; Krueger et al., 2007). Thus, mentalizing and corresponding activity in the frontal cortex, respectively, influences trust behavior.

OXT is also related to the limbic system. OXT is synthesized in the paraventricular nucleus and supraoptic nucleus of the hypothalamus (Zak & Fakhar, 2006). Moreover, limbic regions associated with emotions and memory (e.g., amygdala, hippocampus, and hypothalamus) have an accumulation of OXT receptors (Barberis & Tribollet, 1996; Insel, 1997; Landgraf & Neumann, 2004; Verbalis, 1999). Also, it has been found that OXT acts on limbic regions, in particular the amygdala, by inhibiting excitatory information from these regions to brainstem sites mediating autonomic fear responses (Debiec, 2005; Huber, Pierre, & Ron, 2005; Petrovic, Kalisch, Singer, & Dolan, 2008). In this context, Baumgartner et al. (2008) recently wrote: "[OXT] reduces fear responses during the trust game by reducing activation in the amygdala and connected brainstem effector sites, which in turn enhances subjects' ability to trust in situations characterized by the risk of betrayal" (p. 645).

In addition to brain structures associated with the processing of negative emotions like fear (e.g., amygdala), as well as uncertainty and risk (e.g., insular cortex), a certain limbic region has been shown to process cognitive conflict, namely the ACC (Botvinick et al., 1999, 2004). Because emotion-related brain structures like the amygdala have abundant connections to the ACC, it can be hypothesized that automatic and typically unconscious information processing, in particular in the amygdala, may be followed by the processing of cognitive conflict in the ACC as a result of balancing the rapid visceral perceptions with more deliberate thoughts (Riedl et al., 2010).

Our review reveals a crucial role for the amygdala and hippocampus (plus parahippocampus gyrus) in trust decisions across all three experimental paradigms (see the asterisks in Table 1). Both limbic structures are important for emotion perception and processing. However, in addition to this general role in emotion perception and processing, both the amygdala and hippocampus were also found to be specifically associated with emotion-based memory (Gazzaniga, Ivry, & Mangun, 2009). Although the exact neural mechanisms are not yet fully understood, a basic understanding of the role of the amygdala and hippocampus in emotion-based memory does exist. We describe this mechanism in the following.

Because emotion is a "fuzzy" construct (Phelps & La Bar, 2006, p. 422), brain research has begun to anatomize it. One prominent categorization is based on the factors *valence* (pleasant-unpleasant or good-bad) and *arousal* (the intensity of the internal emotional response, high-low). By using this categorization, scholars can achieve "a more concrete assessment of the emotional reactions elicited by stimuli" (Gazzaniga et al., 2009, p. 367). With respect to memory, established categorizations refer to *duration* (sensory memory: milliseconds to seconds; short-term memory: seconds to minutes; long-term memory: hours to years) and *content* (declarative memory: knowledge to which we have conscious access such as the recall of events in our lives; nondeclarative memory: knowledge to which we do not have conscious access such as motor and cognitive skills; Gazzaniga et al., 2009).

Research has found that amygdala activity is typically associated with negatively valenced emotions (in particular fear) and with high levels of arousal (Phelps & LeDoux, 2005). Moreover, it has been found that the amygdala plays a crucial role in the neural implementation of declarative memory, independent of how a particular stimulus in the world has been linked to potentially aversive consequences (i.e., via instruction, observation, or experience). Specifically, it has also been found that information storage in the hippocampus is mediated by the amygdala (Gazzaniga et al., 2009), and arousal, rather than the valence of emotion, has been demonstrated to mediate memory (McGaugh, 2004). Thus, arousing events are not forgotten as quickly as nonarousing events (Kleinsmith & Kaplan, 1963).

Against this background, the results presented in Table 1 can be interpreted as follows: Independently of the specific trust situation (i.e., playing the trust game, as well as viewing human faces or Web sites with varying degrees of trustworthiness), the amygdala processes the arousal associated with a particular stimulus, thereby mediating the storage of information concerning the trustworthiness of a stimulus in the hippocampus. A high level of arousal at the storage of a stimulus, both with positive and negative valence, positively affects the capacity for remembering it in a future trust situation.

Figure 5 also shows an interplay between the limbic system and COR. It is an established fact that social stress leads to increased amygdala-HPA activation, which in turn elevates COR levels (Takahashi et al., 2005). Moreover, COR affects the limbic system because it influences hippocampus activity. Specifically, if levels of arousal and stress become too high, resulting in pronounced COR elevation, it is possible that social memory becomes impaired because glucocorticoid receptor (GR, Type II receptor)-mediated neuronal pathways result in suppressed synaptic potentiation (Takahashi, 2005). Thus, the relation between COR and synaptic plasticity (which is important for memory) is hypothesized to have an inverted-U shape (de Kloet et al., 1999; Kirschbaum & Hellhammer, 1994).

With respect to the relation between OXT and COR (stress), it was found that intranasally administered OXT significantly reduces salivary COR levels (Ditzen et al., 2009). In line with this finding, another study (Takahashi et al., 2005) reports a negative correlation between interpersonal trust and COR. That is, subjects with higher degrees of interpersonal trust have lower levels of neuroendocrine response (i.e., COR) to social stress (measured with the Trier Social Stress Test). Additional support for these findings comes from research which reports that subjects with social phobia and anxiety disorders, who typically have an exaggerated HPA response to social stressors, exhibit high levels of interpersonal distrust (Furlan, DeMartinis, Schweizer, Rickels, & Lucki, 2001; Takahashi et al., 2005).

We have already outlined that high levels of AVP typically result in distrust behavior (Heinrichs & Domes, 2008; Heinrichs et al., 2009). A possible pathway for this effect could exist via the limbic system (Debiec, 2005). One study (Murgatroyd et al., 2004), for example, has shown that elevated AVP expression in the hypothalamic paraventricular nucleus is associated with increased anxiety levels, which in turn are related to amygdala activity. In line with this finding, it is reported that AVP (a) is correlated with increased vigilance, anxiety, arousal, and activation (Heinrichs & Domes, 2008), and (b) enhances the encoding of both happy and angry social information to make this more memorable (Guastella, Kenyon, Alvares, Carson, & Hickie, 2010). Because all these four states, as well as memory, are related to activity in limbic structures, this finding further supports the notion of an interplay between AVP and limbic brain regions. Finally, because both COR and AVP negatively affect trust behavior, it is no surprise that research has also found a direct positive relation between these two hormones. Specifically, intranasal administration of AVP has been shown to increase salivary COR levels in a social stress test (de Winter et al., 2003; Ebstein et al., 2009; Patchev & Almeida, 1995).

The final hormone which we address in our conceptual model (see Figure 5) is SER. It has been shown that SER depletion leads to a higher rejection rate of unfair offers in a two-person economic game (Crockett et al., 2008), indicating the importance of SER for the control of negative emotions that are processed and controlled both in the limbic system and in the frontal cortex (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Importantly, such negative emotions play a significant role in trust situations (Winston et al., 2002). In line with this result, research (Wood et al., 2006) has found that people with lower SER levels in the brain are less cooperative in economic game playing than control subjects. Moreover, a negative correlation between SER and stress is likely to exist because research has found that touch is associated with increased levels of SER (Field et al., 1996), which have been demonstrated to reduce stress reactivity in humans (Hanley & Van de Kar, 2003). Finally, it is reported that SER desensitizes the AVP receptor (Skuse & Gallagher, 2011), which could reduce approach and trust behavior.

4. Summary, Implications, and Conclusion

In the present article, we have reviewed the literature on the biology of trust. We structured our discussion along a model, which integrates three biological levels of analysis: genes, hormones, and the brain (see Figure 2). Altogether, our review indicates that at least a moderate degree of human trust behavior is genetically predetermined, and several hormones (in particular OXT, see Figure 4), as well as specific brain areas (which are mainly located in the basal ganglia, the limbic system, and the frontal cortex, see Table 1), are strongly associated with human trust behavior. Importantly, a major contribution of the present article is the development of a conceptual framework which unifies the levels of analysis (see Figure 5).

Based on the results of our review, possible avenues for future trust research have emerged. These implications for future research can be grouped into two classes: methodological and thematic.

From a methodological perspective, one major conclusion that can be drawn from our review is that most studies have so far addressed only one level of analysis. Despite the value of such investigations, it is clear that studies integrating two or even all three biological levels of analysis could result in a deeper understanding of the complex relations among the biological determinants of trust behavior. One study (Baumgartner et al., 2008), for example, has already combined administration of OXT and fMRI scanning, thereby triangulating across two different levels of analysis (hormone and brain). As a consequence, this study revealed intriguing insights into the interplay between OXT and brain areas in the basal ganglia (striatum) and limbic system (amygdala). Drawing upon this pioneering study, we call for more investigations that combine the various levels of analysis.

On the brain level of analysis, we chose to focus on functional brain imaging studies, because these studies dominate the literature in terms of quantity. Other methods (e.g., electroencephalography, EEG), in contrast, have been used to a much smaller extent to reveal the neurobiological mechanisms underlying trust. Against this background, future trust research could shift its focus to EEG and other neuroscience (e.g., diffusion tensor imaging) and neurophysiological (e.g., galvanic skin response) tools.

One study (Boudreau, McCubbins, & Coulson, 2008), for example, already investigated human trust behavior by means of EEG. This study signifies the potential of the complementary use of different neuroscience tools to shed light on the neurobiological mechanisms underlying human trust behavior. EEG enables a temporal resolution of milliseconds and this tool can, therefore, easily detect the time course of neural activity. However, since its spatial resolution is limited, triangulation with other tools is essential—fMRI studies, in particular, have already revealed many insights into the neural mechanisms of trust behavior. However, a major implication of our review is that future investigations could use tools which are currently not well represented in studies on the biological foundations of trust (e.g., EEG). Accordingly, research results should be embraced only after they are corroborated by more than one method.

In addition to these methodological implications, our review also reveals thematic avenues for future studies, of which we address three important ones in the following.

We discussed that increasingly more interactions among humans are taking place online. Thus, traditional face-to-face interactions are declining, while computer-mediated interactions are increasing. Two recent fMRI studies (Dimoka, 2010; Riedl et al., 2010) have already addressed this development and investigated the neural correlates of trust while the participants viewed eBay Web sites. On the Internet, however, an increasing number of users have started to represent themselves as computer generated virtual characters, so-called *avatars* (Bainbridge, 2007).[3] Thus, the question arises whether the neurobiological mechanisms underlying human trust behavior in traditional face-to-face settings resemble those in avatar-based online environments.

In one fMRI study (Moser et al., 2007), participants performed facial emotion recognition tasks based on human and avatar stimuli. Although the neural responses were significantly stronger to human faces in face-sensitive structures (in particular in the fusiform gyrus), robust amygdala activation was found in response to both human and avatar emotional faces. Other studies using different techniques to capture neurobiological reactions to avatars and avatar-like robots (heart rate, electrodermal activity, electromyographic activity, and electroencephalography) found similar results (Dubal, Foucher, Jouvent, & Nadel, 2011; Garau, Slater, Pertaub, & Razzaque, 2005; Weyers, Mühlberger, Hefele, & Pauli, 2006). Considering this, current research suggests that avatars have the potential to elicit strong emotional reactions in humans, one of the major foundations for trust, as well as distrust, to emerge (Damasio, 2005; Winston et al., 2002; Zak, 2003). Building on these existing studies, we call for bio-

[3] For example, Second Life, one of the worldwide largest virtual worlds in which users interact with each other through avatars, has currently approximately 24 million user accounts (see www.secondlife.com, September 2011).

logical research with the purpose of revealing the differences and similarities in the biology of trust in traditional face-to-face settings and computer-mediated (avatar-based) environments. Such research efforts will not only reveal theoretical insights into the neurobiology of trust, but also indicate findings which might be relevant for policymakers.

Another finding of our review is that research has focused on the trustor rather than the trustee (see Figure 1). Thus, existing knowledge on the biology of trust and corresponding behavior mainly pertains to one specific party in a trust situation—exceptions are, for example, Cesarini et al. (2008), Zak et al. (2004, 2005b), and King-Casas et al. (2005). Considering this research gap, we call for more investigations that focus on the trustee, as well as the direct interaction between the trustor and trustee (e.g., via hyperscanning).

A final conclusion of our review is that to date, moderator variables such as gender and age have hardly been addressed in research on the biology of trust. With respect to gender, one fMRI study (Riedl et al., 2010) found brain activation differences between men and women when processing trustworthy and untrustworthy eBay offers. These results indicate that well-known gender differences in computer and Internet behavior are correlated with differences in brain functionality. Moreover, behavioral research (e.g., Sutter & Kocher, 2007) found that trust increases almost linearly from early childhood to adulthood. Because the human brain develops across the entire life span, it would be worthwhile to see what insight future research might reveal as to how these behavioral changes are associated with neurobiological changes.

In our concluding statement, we would like to direct the readers' attention to the fact that this review is focused on the *biology* of trust. We have already pointed out that trust behavior is not only influenced by such biological factors, rather, environmental factors, in particular those related to the *social dimension of interaction among humans* (e.g., socialization and culture), also affect human trust behavior (see Figure 2). In this context, Cacioppo et al. (2000, p. 829) write: "All human behavior, at some level, is biological, but this is not to say that biological reductionism yields a simple, singular, or satisfactory explanation for complex behaviors or that molecular forms of representation provide the only or best level of analysis for understanding human behavior . . . constructs such as those developed by the social sciences provide a means of understanding highly complex activity without needing to specify each individual action of the simplest components, thereby offering an efficient means of describing the behavior of a complex system." Against this background, we argue that in order to obtain a comprehensive understanding of human trust behavior, it would be advantageous to integrate the biological and the social perspectives in future research.

References

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature, 433,* 68–72. doi:10.1038/nature03086

Adolphs, R., Tranel, D., & Damasio, A. (1998). The human amygdala in social judgement. *Nature, 393,* 470–474. doi:10.1038/30982

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion Neurobiology, 12,* 169–177. doi:10.1016/S0959-4388(02)00301-X

Aimone, J. A., & Houser, D. (2011). Beneficial betrayal aversion. *PLoS ONE, 6,* e17725. doi:10.1371/journal.pone.0017725

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 268–277. doi:10.1038/nrn1884

Apicella, C. L., Cesarini, D., Johannesson, M., Dawes, C. T., Lichtenstein, P., Wallace, B., . . . Westberg, L. (2010). No association between oxytocin receptor (OXTR) gene polymorphisms and experimentally elicited social preferences. *PLoS ONE, 5,* e11153. doi:10.1371/journal.pone.0011153

Arsenijevic, Y., & Tribollet, E. (1998). Region-specific effect of testosterone on oxytocin receptor binding in the brain of the aged rat. *Brain Research, 785,* 167–170. doi:10.1016/S0006-8993(97)01429-7

Bainbridge, W. S. (2007). The scientific research potential of virtual worlds. *Science, 317,* 472–476. doi:10.1126/science.1146930

Barberis, C., & Tribollet, E. (1996). Vasopressin and oxytocin receptors in the central nervous system. *Critical Reviews in Neurobiology, 10,* 119–154.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21,* 37–46. doi:10.1016/0010-0277(85)90022-8

Barraza, J. A., & Zak, P. J. (2009). Empathy toward strangers triggers oxytocin release and subsequent generosity. *Annals of the New York Academy of Sci-*

ences, 1167, 182–189. doi:10.1111/j.1749-6632.2009.04504.x

Barrot, M., Marinelli, M., Abrous, D. N., Rouge-Pont, F., Le Moal, M., & Piazza, P. V. (2000). The dopaminergic hyper-responsiveness of the shell of the nucleus accumbens is hormone-dependent. *European Journal of Neuroscience, 12,* 973–979. doi:10.1046/j.1460-9568.2000.00996.x

Baskerville, T. A., & Douglas, A. J. (2010). Dopamine and oxytocin interactions underlying behaviors: Potential contributions to behavioral disorders. *CNS Neuroscience & Therapeutics, 16,* e92–e123. doi:10.1111/j.1755-5949.2010.00154.x

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58,* 639–650. doi:10.1016/j.neuron.2008.04.009

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social-history. *Games and Economic Behavior, 10,* 122–142. doi:10.1006/game.1995.1027

Bhatt, M., & Camerer, C. F. (2005). Self-referential thinking and equilibrium as states of mind in games: FMRI evidence. *Games and Economic Behavior, 52,* 424–459. doi:10.1016/j.geb.2005.03.007

Bjornskov, C. (2007). Determinants of generalized trust: A cross-county comparison. *Public Choice, 130,* 1–21. doi:10.1007/s11127-006-9069-1

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review, 98,* 294–310. doi:10.1257/aer.98.1.294

Bos, P. A., Terburg, D., & Van Honk, J. (2010). Testosterone decreases trust in socially naïve humans. *Proceedings of the National Academy of Sciences, 107,* 9991–9995. doi:10.1073/pnas.0911700107

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8,* 539–546. doi:10.1016/j.tics.2004.10.003

Botvinick, M. M., Nystrom, L., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature, 402,* 179–181. doi:10.1038/46035

Bouchard, T. J., Jr. (1994). Genes, environment, and personality. *Science, 264,* 1700–1701. doi:10.1126/science.8209250

Boudreau, C., McCubbins, M. D., & Coulson, S. (2008). Knowing when to trust others: An ERP study of decision making after receiving information from unknown people. *Social Cognitive and Affective Neuroscience, 4,* 23–34. doi:10.1093/scan/nsn034

Cacioppo, J. T., Berntson, G. G., Sheridan, J. F., & McClintock, M. K. (2000). Multilevel integrative analyses of human behavior: Social neuroscience and the complementing nature of social and biological approaches. *Psychological Bulletin, 126,* 829–843. doi:10.1037/0033-2909.126.6.829

Camerer, C. F., & Weigelt, K. (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica, 56,* 1–36. doi:10.2307/1911840

Carter, C. S. (2007). Sex differences in oxytocin and vasopressin: Implications for autism spectrum disorders? *Behavioural Brain Research, 176,* 170–186. doi:10.1016/j.bbr.2006.08.025

Casolini, P., Kabbaj, M., Leprat, E., Piazza, P. V., Rouge-Pont, E., Angelucci, L., . . . Maccari, S. (1993). Basal and stress-induced corticosterone secretion is decreased by lesion of mesenchephalic dopaminergic neurons. *Brain Research, 622,* 311–314. doi:10.1016/0006-8993(93)90836-C

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., & Wallace, B. (2008). Heritability of cooperative behavior in the trust game. *Proceedings of the National Academy of Sciences, 105,* 3721–3726. doi:10.1073/pnas.0710069105

Coleman, J. S. (1990). *Foundations of social theory.* Cambridge, MA: Harvard University Press.

Cowen, P. J. (2002). Cortisol, serotonin and depression: All stressed out? *The British Journal of Psychiatry, 180,* 99–100. doi:10.1192/bjp.180.2.99

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron, 29,* 537–545. doi:10.1016/S0896-6273(01)00225-2

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science, 320,* 1739. doi:10.1126/science.1155577

Damasio, A. (2005). Brain trust. *Nature, 435,* 571–572. doi:10.1038/435571a

Debiec, J. (2005). Peptides of love and fear: Vasopressin and oxytocin modulate the integration of information in the amygdala. *BioEssays, 27,* 869–873. doi:10.1002/bies.20301

De Dreu, C. K. W., Greer, L. L., Handgraaf, M. J. J., Shalvi, S., Van Kleef, G. A., Baas, M., . . . Feith, S. W. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science, 328,* 1408–1411. doi:10.1126/science.1189047

de Kloet, E. R., Oitzl, M. S., & Joels, M. (1999). Stress and cognition: Are corticosteroids good or bad guys? *Trends in Neuroscience, 22,* 422–426. doi:10.1016/S0166-2236(99)01438-1

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8,* 1611–1618. doi:10.1038/nn1575

de Winter, R. F., Van Hemert, A. M., DeRijk, R. H., Zwinderman, K. H., Frankhuijzen-Sierevogel, A. C., Wiegant, V. M., & Goekopp, J. G. (2003). Anxious-retarded depression: Relation with plasma vasopressin and cortisol. *Neuropsychopharmacology, 28,* 140–147. doi:10.1038/sj.npp.1300002

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly, 34,* 373–396.

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological Psychiatry, 65,* 728–731. doi:10.1016/j.biopsych.2008.10.011

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2006). The intergenerational transmission of risk and trust attitudes. IZA Discussion Paper, No. 2380, The Institute for the Study of Labor, Bonn, Germany.

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. (2007). Oxytocin improves "mind-reading" in humans. *Biological Psychiatry, 61,* 731–733. doi:10.1016/j.biopsych.2006.07.015

Donaldson, Z. R., & Young, L. J. (2008). Oxytocin vasopressin and the neurogenetics of sociality. *Science, 322,* 900–904. doi:10.1126/science.1158668

Dubal, S., Foucher, A., Jouvent, R., & Nadel, J. (2011). Human brain spots emotion in non humanoid robots. *Social Cognitive and Affective Neuroscience, 6,* 90–97. doi:10.1093/scan/nsq019

Ebstein, R. P., Israel, S., Lerer, E., Uzefovsky, F., Shalev, I., Gritsenko, I., . . . Yirmiya, N. (2009). Arginine vasopressin and oxytocin modulate human social behavior. *Annals of the New York Academy of Sciences, 1167,* 87–102. doi:10.1111/j.1749-6632.2009.04541.x

Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., & Fehr, E. (2010). Prejudice and truth about the effect of testosterone on human bargaining behaviour. *Nature, 463,* 356–359. doi:10.1038/nature08711

Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *American Economic Review, 96,* 1611–1630. doi:10.1257/aer.96.5.1611

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences, 11,* 419–427. doi:10.1016/j.tics.2007.09.002

Fehr, E. (2009a). On the economics and biology of trust. *Journal of the European Economic Association, 7,* 235–266. doi:10.1162/JEEA.2009.7.2-3.235

Fehr, E. (2009b). Social preferences and the brain. In P. W. Glimcher, C. F. Camerer, E. Fehr, and R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain,* (pp. 215–232). Amsterdam, The Netherlands: Elsevier.

Fehr, E. Fischbacher, U., & Kosfeld, M. (2005). Neuroeconomic foundations of trust and social preferences: Initial evidence. *American Economic Review, 95,* 346–351. doi:10.1257/000282805774669736

Ferris, C. F., & Delville, Y. (1994). Vasopressin and serotonin interactions in the control of agonistic behavior. *Psychoneuroendocrinology, 19,* 593–601. doi:10.1016/0306-4530(94)90043-4

Ferris, C. F., Melloni, Jr., R. H., Koppel, G., Perry, K. W., Fuller, R. W., & Delville, Y. (1997). Vasopressin/serotonin interactions in the anterior hypothalamus control aggressive behavior in golden hamsters. *Journal of Neuroscience, 17,* 4331–4340.

Field, T., Grizzle, N., Scafidi, F., & Schanberg, S. (1996). Massage and relaxation therapies' effects on depressed adolescent mothers. *Adolescence, 31,* 903–911.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50,* 531–534. doi:10.1016/j.neuron.2006.05.001

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London–B, 358,* 459–473. doi:10.1098/rstb.2002.1218

Frith, U., & Frith, C. D. (2010). The social brain: Allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society of London–B, 365,* 165–176. doi:10.1098/rstb.2009.0160

Füllbrunn, S., Richwien, K., & Sadrieh, A. (2011). Trust and trustworthiness in anonymous virtual worlds. *Journal of Media Economics, 24,* 48–63. doi:10.1080/08997764.2011.549432

Furlan, P. M., DeMartinis, N., Schweizer, E., Rickels, K., & Lucki, I. (2001). Abnormal salivary cortisol levels in social phobic patients in response to acute psychological but not physical stress. *Biological Psychiatry, 50,* 254–259. doi:10.1016/S0006-3223(00)01126-4

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage, 16,* 814–821. doi:10.1006/nimg.2002.1117

Garau, M., Slater, M., Pertaub, D.-P., & Razzaque, S. (2005). The responses of people to virtual humans in an immersive virtual environment. *Presence, 14,* 104–116. doi:10.1162/1054746053890242

Gazzaniga, M. S., Ivry, R., & Mangun, G. R. (2009). *Cognitive neuroscience: The biology of the mind* (3rd edition), New York, NY: W. W. Norton.

Gefen, D., Benbasat, I., & Pavlou, P. A. (2008). A research agenda for trust in online environments. *Journal of Management Information Systems, 24,* 275–286. doi:10.2753/MIS0742–1222240411

Guastella, A. J., Kenyon, A. R., Alvares, G. A., Carson, D. S., & Hickie, I. B. (2010). Intranasal arginine vasopressin enhances the encoding of happy and angry faces in humans. *Biological Psychiatry, 67,* 1220–1222. doi:10.1016/j.biopsych.2010.03.014

Hanley, N. R., & Van de Kar, L. D. (2003). Serotonin and the neuroendocrine regulation of the hypothalamic-pituitary-adrenal axis in health and disease. *Vitamins and Hormones, 66,* 189–255. doi:10.1016/S0083-6729(03)01006-9

Harlan, R. E. (1988). Regulation of neuropeptide gene expression by steroid hormones. *Molecular Neurobiology, 2,* 183–200. doi:10.1007/BF02935345

Harris, G. W. (1948). Neural control of the pituitary gland. *Physiological Reviews, 28,* 139–179.

Harris, J. A., Vernon, P. A., & Boomsma, D. I. (1998). The heritability of testosterone: A study of Dutch adolescent twins and their parents. *Behavior Genetics, 28,* 165–171. doi:10.1023/A:1021466929053

Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., & Rauch, S. L. (2000). Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *Neuroreport, 11,* 2351–2355. doi:10.1097/00001756-200008030-00004

Heinrichs, M., Baumgartner, T., Kirschbaum, C., & Ehlert, U. (2003). Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. *Biological Psychiatry, 54,* 1389–1398. doi:10.1016/S0006-3223(03)00465-7

Heinrichs, M., & Domes, G. (2008). Neuropeptides and social behavior: Effects of oxytocin and vasopressin in humans. *Progress in Brain Research, 170,* 337–350. doi:10.1016/S0079-6123(08)00428-7

Heinrichs, M., Meinlschmidt, G., Wippich, W., Ehlert, U., & Hellhammer, D. H. (2004). Selective amnesic effects of oxytocin on human memory. *Physiology & Behavior, 83,* 31–38.

Heinrichs, M., von Dawans, B., & Domes, G. (2009). Oxytocin, vasopressin, and human social behavior. *Frontiers in Neuroendocrinology, 30,* 548–557. doi:10.1016/j.yfrne.2009.05.005

Huber, D., Pierre, V., & Ron, S. (2005). Vasopressin and oxytocin excite distinct neuronal populations in the central amygdala. *Science, 308,* 245–248. doi:10.1126/science.1105636

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., & Platt, M. L. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron, 49,* 765–775. doi:10.1016/j.neuron.2006.01.024

Ikemoto, S., & Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: A unifying interpretation with special reference to reward-seeking. *Brain Research Reviews, 31,* 6–41. doi:10.1016/S0165-0173(99)00023-5

Inoue, T., Kimura, T., Azuma, C., Inazawa, J., Takemura, M., Kikuchi, T., . . . Saji, F. (1994). Structural organization of the human oxytocin receptor gene. *Journal of Biological Chemistry, 269,* 32451–32456.

Insel, T. R. (1997). The neurobiology of social attachment. *American Journal of Psychiatry, 154,* 726–735.

Insel, T. R., & Young, L. J. (2001). The neurobiology of attachment. *Nature Reviews Neuroscience, 2,* 129–136. doi:10.1038/35053579

Internet World Stats (2010). Retrieved February 21, 2011 from http://www.internetworldstats.com

Israel, S., Lerer, E., Shalev, I., Uzefovsky, F., Riebold, M., Laiba, E., . . . Ebstein, R. P. (2009). The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task. *PLoS ONE, 4,* e5535. doi:10.1371/journal.pone.0005535

Johnson, R. T., & Breedlove, S. M. (2010). Human trust: Testosterone raises suspicion. *Proceedings of the National Academy of Sciences, 107,* 11149–11150. doi:10.1073/pnas.1005376107

Johnson, W. (2007). Genetic and environmental influences on behavior: Capturing all the interplay. *Psychological Review, 114,* 423–440. doi:10.1037/0033-295X.114.2.423

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308,* 78–83. doi:10.1126/science.1108062

Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., . . . Meyer-Lindenberg, A. (2005). Oxytocin modulates neural circuitry for social cognition and fear in humans. *Journal of Neuroscience, 25,* 11489–11493. doi:10.1523/JNEUROSCI.3984-05.2005

Kirschbaum, C., & Hellhammer, D. H. (1994). Salivary cortisol in psychoneuroendocrine research: Recent developments and applications. *Psychoneuroendocrinology, 19,* 313–333. doi:10.1016/0306-4530(94)90013-2

Kleinsmith, L. J., & Kaplan, S. (1963). Paired-associate learning as a function of arousal and interpolated interval. *Journal of Experimental Psychology, 65,* 190–193. doi:10.1037/h0040288

Knafo, A., Israel, S., Darvasi, A., Bachner-Melman, R., Uzefovsky, F., Cohen, L., . . . Ebstein, R. P. (2008). Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. *Genes Brain & Behavior, 7,* 266–275. doi:10.1111/j.1601-183X.2007.00341.x

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior, 20,* 165–182. doi:10.1007/BF02281954

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., & Ono, T. (2001). Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature, 412,* 546–549. doi:10.1038/35087595

Koscik, T. R., & Tranel, D. (2011). The human amygdala is necessary for developing and expressing normal interpersonal trust. *Neuropsychologia, 49,* 602–611. doi:10.1016/j.neuropsychologia.2010.09.023

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435,* 673–676. doi:10.1038/nature03701

Krain, A. L., Wilson, A. M., Arbuckle, R., Castellanos, F. X., & Milham, M. P. (2006). Distinct neural mechanisms of risk and ambiguity: A meta-analysis of decision-making. *Neuroimage, 32,* 477–484. doi:10.1016/j.neuroimage.2006.02.047

Krueger, F., Grafman, J., & McCabe, K. (2008). Neural correlates of economic game playing. *Philosophical Transactions of the Royal Society of London–B, 363,* 3859–3874. doi:10.1098/rstb.2008.0165

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., . . . Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences, 104,* 20084–20089. doi:10.1073/pnas.0710103104

Kugler, T., Connolly, T., & Kausel, E. E. (2009). The effect of consequential thinking on trust game behavior. *Journal of Behavioral Decision Making, 22,* 101–119. doi:10.1002/bdm.614

Kuhnen, C. M., & Knutson, B. (2005). The neural basis of financial risk-taking. *Neuron, 47,* 763–770. doi:10.1016/j.neuron.2005.08.008

Landgraf, R., & Neumann, I. D. (2004). Vasopressin and oxytocin release within the brain: A dynamic concept of multiple and variable modes of neuropeptide communication. *Frontiers in Neuroendocrinology, 25,* 150–176. doi:10.1016/j.yfrne.2004.05.001

Lee, Y. Y., & Lin, J. L. (2009). The effects of trust in physician on self-efficacy, adherence and diabetes outcomes. *Social Science and Medicine, 68,* 1060–1068. doi:10.1016/j.socscimed.2008.12.033

Lieberman, M. D., & Eisenberger, N. I. (2009). Pains and pleasures of social life. *Science, 323,* 890–891. doi:10.1126/science.1170008

Loup, F., Tribollet, E., Dubois-Dauphin, M., & Dreifuss, J. J. (1991). Localization of high affinity binding sites for oxytocin and vasopressin in the human brain: An autoradiographic study, *Brain Research, 555,* 220–232. doi:10.1016/0006-8993(91)90345-V

Luhmann, N. (1979). *Trust and power*. New York, NY: John Wiley & Sons.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20,* 709–734.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, 98,* 11832–11835. doi:10.1073/pnas.211415698

McCabe, K., Rigdon, M., & Smith, V. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization, 52,* 267–275. doi:10.1016/S0167-2681(03)00003-9

McCabe, K. A., & Smith, V. (2000). A two person trust game played by naive and sophisticated subjects. *Proceedings of the National Academy of Sciences, 97,* 3777–3781. doi:10.1073/pnas.040577397

McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience, 27,* 1–28. doi:10.1146/annurev.neuro.27.070203.144157

McKittrick, C. R., Blanchard, D. C., Hardy, M. P., & Blanchard, R. J. (2009). Social stress effects on hormones, brain, and behaviour. In D. W. Pfaff, A. P. Arnold, A. M. Etgen, S. E. Fahrbach, and R. T. Rubin (Eds.), *Hormones, brain and behavior*. (pp. 341−347). San Diego, CA: Academic. doi:10.1016/B978-008088783-8.00009-7

McKnight, H. D., & Chervany, N. L. (2001). What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International Journal of Electronic Commerce, 6,* 35–59.

McKnight, H. D., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13,* 334–359. doi:10.1287/isre.13.3.334.81

Mikolajczak, M., Gross, J. J., Lane, A., Corneille, O., de Timary, P., & Luminet, O. (2010). Oxytocin makes people trusting, not gullible. *Psychological Science, 21,* 1072–1074. doi:10.1177/0956797610377343

Miller, G. (2005). Economic game shows how the brain builds trust. *Science, 308,* 36. doi:10.1126/science.308.5718.36a

Morris, J., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature, 383,* 812–815. doi:10.1038/383812a0

Moser, E., Derntl, B., Robinson, S., Fink, B., Gur, R. C., & Grammer, K. (2007). Amygdala activation at 3T in response to human and avatar facial expressions of emotions. *Journal of Neuroscience Methods, 161,* 126–133. doi:10.1016/j.jneumeth.2006.10.016

Murgatroyd, C., Wigger, A., Frank, E., Singewald, N., Bunck, M., Holsboer, F., . . . Spengle, D. (2004). Impaired repression at a vasopressin promoter polymorphism underlies overexpression of vasopressin in a rat model of trait anxiety. *Journal of Neuroscience, 24,* 7762–7770. doi:10.1523/JNEUROSCI.1614-04.2004

Nguyen, G. C., Laveist, T. A., Harris, M. L., Datta, L. W., Bayless, T. M., & Brant, S. R. (2009). Patient trust-in-physician and race are predictors of adherence to medical management in inflammatory bowel disease. *Inflammatory Bowel Diseases, 15,* 1233–1239. doi:10.1002/ibd.20883

Numan, R. (2000). *The behavioral neuroscience of the septal region.* New York, NY: Springer. doi:10.1007/978-1-4612-1302-4

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304,* 452–454. doi:10.1126/science.1094285

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105,* 11087–11092. doi:10.1073/pnas.0805664105

Patchev, V. K., & Almeida, O. F. X. (1995). Corticosteroid regulation of gene expression and binding characteristics of vasopressin receptors in the rat brain. *European Journal of Neuroscience, 7,* 1579–1583. doi:10.1111/j.1460-9568.1995.tb01153.x

Pavlou, P. A., Liang, H., & Xue, Y. (2007). Understanding and Mitigating Uncertainty in Online Exchange Relationships: A Principal Agent Perspective. *MIS Quarterly, 31,* 1–32.

Petrovic, P., Kalisch, R., Singer, T., & Dolan, R. J. (2008). Oxytocin attenuates affective evaluations of conditioned faces and amygdala activity. *Journal of Neuroscience, 28,* 6607–6615. doi:10.1523/JNEUROSCI.4572-07.2008

Phelps, E. A., & LaBar, K. S. (2006). Functional neuroimaging of emotion and social cognition. In R. Cabeza & A. Kingstone (Eds.), *Handbook of functional neuroimaging of cognition* (pp. 421–453). Cambridge, MA: MIT Press.

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behaviour. *Neuron, 48,* 175–187. doi:10.1016/j.neuron.2005.09.025

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12,* 729–738. doi:10.1162/089892900562552

Phillips, M., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., . . . Gray, J. A. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society of London – B, 265,* 1809–1817. doi:10.1098/rspb.1998.0506

Phillips, M., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., . . . David, A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389,* 495–498. doi:10.1038/39051

Powell, E. W., & Rorie, D. (1967). Septal projections to nuclei functioning in oxytocin release. *American Journal of Anatomy, 120,* 605–610. doi:10.1002/aja.1001200310

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a "theory of mind"? *Behavioral and Brain Sciences, 1,* 515–526. doi:10.1017/S0140525X00076512

Prewitt, C. M., & Herman, J. P. (1994). Lesion of the central nucleus of the amygdala decreases CRH mRNA expression and stress-induced ACTH release. *Annals of the New York Academy of Sciences, 746,* 438–440. doi:10.1111/j.1749-6632.1994.tb39279.x

Price, C. J., & Friston, K. J. (2005). Functional ontologies for cognition: The systematic definition of structure and function, *Cognitive Neuropsychology, 22,* 262–275. doi:10.1080/02643290442000095

Reuter, M., Montag, C., Altmann, S., Bendlow, F., Elger, C., Kirsch, P., . . . Falk, A. (2009). Genetically determined differences in human trust behavior: The role of the oxytocin receptor gene. *NeuroPsychoEconomics Conference Proceedings,* 21.

Riedl, R., Hubert, M., & Kenning, P. (2010). Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *MIS Quarterly, 34,* 397–428.

Righetti, F., & Finkenauer, C. (2011). If you are able to control yourself, I will trust you: The role of perceived self-control in interpersonal trust. *Journal of Personality and Social Psychology, 100,* 874–886. doi:10.1037/a0021827

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron, 35,* 395–405. doi:10.1016/S0896-6273(02)00755-9

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology, 62,* 23–48. doi:10.1146/annurev.psych.121208.131647

Rimmele, U., Spillmann, M., Bärtschi, C., Wolf, O. T., Weber, C. S., Ehlert, U., & Wirtz, P. H. (2009). Melatonin improves memory acquisition under stress independent of stress hormone release. *Psychopharmacology, 202,* 663–672. doi:10.1007/s00213-008-1344-z

Rodrigues, S. M., Saslow, L. R., Garcia, N., John, O. P., & Keltner, D. (2009). Oxytocin receptor genetic variation relates to empathy and stress reactivity in humans. *Proceedings of the National Academy of Sciences, 106,* 21437–21441. doi:10.1073/pnas.0909579106

Rostene, W., Sarrieau, A., Nicot, A., Scarceriaux, V., Betancur, C., Gully, D., . . . Berod, A. (1995). Steroid effects on brain functions: An example of the action of glucocorticoids on central dopaminergic and neurotensinergic systems. *Journal of Psychiatry & Neuroscience, 20,* 349–356.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. F. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23,* 393–404. doi:10.5465/AMR.1998.926617

Rushton, J. P., & Ankney, C. D. (2007). The evolution of brain size and intelligence. In S. M. Platek, J. P. Keenan, & T. K. Shackelford (Eds.), *Evolutionary cognitive neuroscience,* (pp. 121–161). Cambridge, MA: MIT Press.

Rushton, J. P., & Jensen, A. R. (2008). James Watson's most inconvenient truth: Race realism and the moralistic fallacy. *Medical Hypotheses, 71,* 629–640. doi:10.1016/j.mehy.2008.05.031

Sanchez-Pages, S., & Turiegano, E. (2010). Testosterone, facial symmetry and cooperation in the prisoners' dilemma. *Physiology & Behavior, 99,* 355–361. doi:10.1016/j.physbeh.2009.11.013

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300,* 1755–1758. doi:10.1126/science.1082976

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599. doi:10.1126/science.275.5306.1593

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience, 1,* 199–207. doi:10.1038/35044563

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron, 36,* 241–263. doi:10.1016/S0896-6273(02)00967-4

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology, 57,* 87–115. doi:10.1146/annurev.psych.56.091103.070229

Seppänen, R., Blomqvist, K., & Sundqvist, S. (2007). Measuring inter-organizational trust: A critical review of the empirical research in 1990–2003. *Industrial Marketing Management, 36,* 249–265. doi:10.1016/j.indmarman.2005.09.003

Shahrokh, D. K., Zhang, T. Y., Diorio, J., Gratton, A., & Meaney, M. J. (2010). Oxytocin-dopamine interactions mediate variations in maternal behavior in the rat. *Endocrinology, 151,* 2276–2286. doi:10.1210/en.2009-1271

Siegal, M., & Varley, R. (2002). Neural systems involved in "theory of mind." *Nature Reviews Neuroscience, 3,* 463–471.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron, 41,* 653–662. doi:10.1016/S0896-6273(04)00014-5

Singer, T. (2009). Understanding others: Brain mechanisms of theory of mind and empathy. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 251–268). Amsterdam, The Netherlands: Elsevier.

Skuse, D. H., & Gallagher, L. (2011). Genetic influences on social cognition. *Pediatric Research, 69,* 85R–91R. doi:10.1203/PDR.0b013e318212f562

Smith, J. H. (2010). Trusting the avatar. *Games and Culture, 5,* 298–313. doi:10.1177/1555412009359764

Sturgis, P., Read, S., Hatemi, P. K., Zuh, G., Trull, T., Wright, M. J., & Martin, M. G. (2010). A genetic basis for social trust. *Political Behavior, 32,* 205–230. doi:10.1007/s11109-009-9101-5

Stuss, D. T., Gallup G. G. Jr., & Alexander, M. P. (2001). The frontal lobes are necessary for "theory of mind." *Brain, 124,* 279–286. doi:10.1093/brain/124.2.279

Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior, 59,* 364–382.

Swan, J. E., Bowers, M. R., & Richardson, L. D. (1999). Customer trust in the salesperson: An integrative review and meta-analysis of the empirical literature. *Journal of Business Research, 44,* 93–107. doi:10.1016/S0148-2963(97)00244-0

Takahashi, T., Ikeda, K., Ishikawa, M., Kitamura, N., Tsukasaki, T., Nakama, D., & Kameda, T. (2005). Interpersonal trust and social stress-induced cortisol elevation. *Neuroreport, 16,* 197–199. doi:10.1097/00001756-200502080-00027

Takahashi, T., Ikeda, K., Ishikawa, M., Tsukasaki, T., Nakama, D., Tanida, S., & Kameda, T. (2004). Social stress-induced cortisol elevation acutely impairs social memory in humans. *Neuroscience Letters, 363,* 125–130. doi:10.1016/j.neulet.2004.03.062

Takahashi, T. (2005). Social memory, social stress, and economic behaviors. *Brain Research Bulletin, 67,* 398–402. doi:10.1016/j.brainresbull.2005.06.006

Taylor, S. E., Burklund, L. J., Eisenberger, N. I., Lehman, B. J., Hilmert, C. J., & Lieberman, M. D. (2008). Neural bases of moderation of cortisol stress responses by psychosocial resources. *Journal of Personality and Social Psychology, 95,* 197–211. doi:10.1037/0022-3514.95.1.197

Theodoridou, A., Rowe, A. C., Penton-Voak, I. S., & Rogers, P. J. (2009). Oxytocin and social perception: Oxytocin increases perceived facial trustworthiness and attractiveness. *Hormones and Behavior, 56,* 128–132. doi:10.1016/j.yhbeh.2009.03.019

Thompson, R. R., George, K., Walton, J. C., Orr, S. P., & Benson, J. (2006). Sex-specific influences of vasopressin on human social communication. *Proceedings of the National Academy of Sciences, 103,* 7889–7894. doi:10.1073/pnas.0600406103

Thompson, R., Gupta, S., Miller, K., Mills, S., & Orr, S. (2004). The effects of vasopressin on human facial responses related to social communication. *Psychoneuroendocrinology, 29,* 35–48. doi:10.1016/S0306-4530(02)00133-6

Tillfors, M., Furmark, T., Marteinsdottir, I., Fischer, H., Pissiota, A., Langstrom, B., & Fredrikson, M. (2001). Cerebral blood flow in subjects with social phobia during stressful speaking tasks: A PET study. *American Journal of Psychiatry, 158,* 1220–1226. doi:10.1176/appi.ajp.158.8.1220

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience, 3,* 119–127. doi:10.1093/scan/nsn009

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/ avoidance behaviors. *Annals of the New York Academy of Sciences, 1124,* 208–224. doi:10.1196/annals.1440.012

Tost, H., Kolachana, B., Hakimi, S., Lemaitre, H., Verchinski, B. A., Mattay, V. S., . . . Meyer-Lindenberg, A. (2010). A common allele in the oxytocin receptor gene (OXTR) impacts prosocial temperament and human hypothalamic-limbic structure and function. *Proceedings of the National Academy of Sciences, 107,* 13936–13941. doi:10.1073/pnas.1003296107

Toulmin, S. E. (1958). *The use of arguments.* Cambridge, MA: Cambridge University Press.

Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron, 41,* 281–292. doi:10.1016/S0896-6273(03)00848-1

Turkheimer, E. (1998). Heritability and biological explanation, *Psychological Review, 105,* 782–791. doi:10.1037/0033-295X.105.4.782-791

Verbalis, J. (1999). The brain oxytocin receptor(s?). *Frontiers in Neuroendocrinology, 20,* 146–156. doi:10.1006/frne.1999.0178

Vilares, I., Dam, G., & Kording, K. (2011). Trust and reciprocity: Are effort and money equivalent? *PLoS ONE, 6,* e17113. doi:10.1371/journal.pone.0017113

Welch, M. R., Rivera, R., Conway, B. P., Yonkoski, J., Lupton, P. M., & Giancola, R. (2005). Determinants and consequences of social trust. *Sociological Inquiry, 75,* 453–473. doi:10.1111/j.1475-682X.2005.00132.x

Wermter, A. K., Kamp-Becker, I., Hesse, P., Schulte-KÖrne, G., Strauch, K., & Remschmidt, H. (2010). Evidence for the involvement of genetic variation in the oxytocin receptor gene (OXTR) in the etiology of autistic disorders on high functioning level. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics, 153B,* 629–639.

Weyers, P., Mühlberger, A., Hefele, C., & Pauli, P. (2006). Electromyographic responses to static and dynamic avatar emotional facial expressions. *Psychophysiology, 43,* 450–453. doi:10.1111/j.1469-8986.2006.00451.x

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5,* 277–283. doi:10.1038/nn816

Wood, R. M., Rilling, J. K., Sanfey, A. G., Bhagwagar, Z., & Rogers, R. D. (2006). Effects of tryptophan depletion on the performance of an iterated prisoner's dilemma game in healthy adults. *Neuropsychopharmacology, 31,* 1075–1084. doi:10.1038/sj.npp.1300932

Yanagisawa, K., Masui, K., Furutani, K., Nomura, M., Ura, M., & Yoshida, H. (2011). Does higher general trust serve as a psychosocial buffer against social pain? An NIRS study of social exclusion. *Social Neuroscience, 6,* 190–197. doi:10.1080/17470919.2010.506139

Zak, P. J., Borja, K., Matzner, W. T., & Kurzban, R. (2005a). The neuroeconomics of distrust: Sex differences in behavior and physiology. *American Economic Review, 95,* 360–363. doi:10.1257/000282805774669709

Zak, P. J., & Fakhar, A. (2006). Neuroactive hormones and interpersonal trust: International evidence. *Economics and Human Biology, 4,* 412–429. doi:10.1016/j.ehb.2006.06.004

Zak, P. J., & Knack, S. (2001). Trust and growth. *The Economic Journal, 111,* 295–321. doi:10.1111/1468-0297.00609

Zak, P. J., Kurzban, R., Ahmadi, S., Swerdloff, R. S., Park, J., Efremidze, L., . . . Matzner, W. T. (2009). Testosterone administration decreases generosity in the ultimatum game. *PLoS ONE, 4,* e8330. doi:10.1371/journal.pone.0008330

Zak, P. J., Kurzban, R., & Matzner, W. T. (2004). The neurobiology of trust. *Annals of the New York Academy of Sciences, 1032,* 224–227. doi:10.1196/annals.1314.025

Zak, P. J., Kurzban, R., & Matzner, W. T. (2005b). Oxytocin is associated with human trustworthiness. *Hormones and Behavior, 48,* 522–527. doi:10.1016/j.yhbeh.2005.07.009

Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS ONE, 2,* e1128. doi:10.1371/journal.pone.0001128

Zak, P. J. (2003). Trust. *Capco Institute Journal of Financial Transformation, 7,* 13–21.

Zak, P. J. (2011). The physiology of moral sentiments. *Journal of Economic Behavior & Organization, 77,* 53–65. doi:10.1016/j.jebo.2009.11.009

Zethraeus, N., Kocoska-Maras, L., Ellingsen, T., von Schoultz, B., Hirschberg, A., & Johannesson, M. (2009). A randomized trial of the effect of estrogen and testosterone on economic behaviour. *Proceedings of the National Academy of Sciences, 106,* 6535–6538. doi:10.1073/pnas.0812757106

Received June 10, 2011
Revision received September 19, 2011
Accepted October 13, 2011 ■

Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory

RENÉ RIEDL, PETER N.C. MOHR, PETER H. KENNING, FRED D. DAVIS, AND HAUKE R. HEEKEREN

René Riedl is a professor of digital business and innovation at the University of Applied Sciences Upper Austria and an associate professor of business informatics at the University of Linz. He was a guest researcher and guest professor at universities in the United States, Germany, and Canada. He earned his Ph.D. from the University of Linz. His research interests include NeuroIS, human–computer interaction, strategic IT management, and technostress. His work has been published in various journals, including *Advances in Human–Computer Interaction, Behavior Research Methods, Business & Information Systems Engineering, DATA BASE for Advances in Information Systems, Journal of Computer Information Systems, Journal of Information Technology, Journal of Neuroscience, Psychology, and Economics,* and *MIS Quarterly.* Moreover, he has editorial roles in several scientific journals, and he also serves as an ad hoc reviewer for the National Science Foundation, the German Research Foundation (DFG), and for several leading journals.

Peter N.C. Mohr is an assistant professor of information processing and economic decision making at the University of Constance, Germany. He studied business administration in Münster, Germany, and received his Ph.D. in psychology from Freie Universität Berlin. His research interests include situational and contextual influences on economic decision making as well as individual differences in this domain. He focuses on behavioral modeling of decision processes as well as on their neural underpinnings.

Peter H. Kenning is a professor of marketing at Zeppelin University, Germany. He received his Ph.D. in business administration from the University of Münster, Germany. His research interests include consumer behavior, consumer neuroscience, neuroeconomics, and marketing management. His work has been published widely, including in *International Journal of Research in Marketing, MIS Quarterly, Journal of Economic Psychology, Journal of Behavioral Decision Making, Journal of Neuroimaging, NeuroReport,* and the *Brain Research Bulletin.* He received several best-paper awards and grants. Since 2011 he has served as a member of the Scientific Advisory Board on Consumer and Food Policies at the Federal Ministry of Consumer Protection, Food and Agriculture (BMELV) in Germany.

Fred D. Davis is Distinguished Professor and David D. Glass Chair in Information Systems at the Sam M. Walton College of Business, University of Arkansas. He earned his Ph.D. from MIT's Sloan School of Management. He served on the Business School faculties of the University of Michigan, University of Minnesota, and University of Maryland. His research interests include user acceptance of information technology, technology-supported decision making, system development practices, and NeuroIS.

Journal of Management Information Systems / Spring 2014, Vol. 30, No. 4, pp. 83–113.

ISSN 0742–1222 (print) / ISSN 1557–928X (online)
DOI: 10.2753/MIS0742-1222300404

His research has been published in *MIS Quarterly, Information Systems Research, Management Science, Journal of Applied Psychology,* and other leading journals. He is currently a senior editor for *Information Systems Research.*

Hauke R. Heekeren is a professor of affective neuroscience and psychology of emotion at the Cluster of Excellence "Languages of Emotion" at Freie Universität Berlin. He received his Ph.D. in medicine from the Charité University of Medicine Berlin. His primary research interests concern the roles of motivation and affect in decision making, cognitive and affective components in normal and disturbed social cognition, and multimodal neuroimaging methodology. Current themes of his research include the effect of neuromodulators (such as drugs or common genetic polymorphisms) on cognitive and affective functions. His work has been published in various outlets, including *Nature, Nature Reviews Neuroscience, Journal of Neuroscience, PLOS Biology, Current Biology,* and *Proceedings of the National Academy of Sciences.*

Abstract: Avatars, as virtual humans possessing realistic faces, are increasingly used for social and economic interaction on the Internet. Research has already determined that avatar-based communication may increase perceived interpersonal trust in anonymous online environments. Despite this trust-inducing potential of avatars, however, we hypothesize that in trust situations, people will perceive human faces differently than they will perceive avatar faces. This prediction is based on evolution theory, because throughout human history the majority of interaction among people has taken place in face-to-face settings. Therefore, unlike perception of an avatar face, perception of a human face and the related trustworthiness discrimination abilities must be part of the genetic makeup of humans. Against this background, we conducted a functional magnetic resonance imaging experiment based on a multiround trust game to gain insight into the differences and similarities of interactions between humans versus human interaction with avatars. Our results indicate that (1) people are better able to predict the trustworthiness of humans than the trustworthiness of avatars; (2) decision making about whether or not to trust another actor activates the medial frontal cortex significantly more during interaction with humans, if compared to interaction with avatars; this brain area is of paramount importance for the prediction of other individuals' thoughts and intentions (mentalizing), a notably important ability in trust situations; and (3) the trustworthiness learning rate is similar, whether interacting with humans or avatars. Thus, the major implication of this study is that although interaction on the Internet may have benefits, the lack of real human faces in communication may serve to reduce these benefits, in turn leading to reduced levels of collaboration effectiveness.

Key words and phrases: agent, avatar, brain, cognitive neuroscience, evolutionary psychology, evolution theory, functional magnetic resonance imaging (fMRI), medial frontal cortex (MFC), mentalizing, NeuroIS, theory-of-mind (TOM).

Determining who can be trusted and who cannot has been crucial since the days of ancient civilizations. From the time of the emergence of early hominids such as *Australopithecus afarensis,* some 3.5 million years ago, until the recent past, trust in individuals who turned out not to be trustworthy could result in loss, or even death [21, 68]. Thus, it is critical for humans to possess distinctive trustworthiness prediction abilities. Attention to another person's face in a trust situation is of particular importance because the human face has been shown to allow one individual to predict another's emotional state, intention, and behavior (e.g., [33, 57]). This predictive function has demonstrated application to trustworthiness [116, 121, 122]. Thus, all other things being equal, the inability to see another person's face creates more difficulty for predicting the other individual's trustworthiness than there would be in a situation in which facial information is available.

Biological anthropologists (e.g., [14, 20]) report that face-to-face interaction has been the main communication medium for more than 99 percent of human history (i.e., since the emergence of the first hominids). It follows, then, that application of Darwin's theory of evolution [23] would suggest that the human trait of processing facial information, and the related prediction abilities, must be part of the genetic makeup of humans—a notion that is widely supported in the scientific literature [21, 23, 121]. For example, Kock and Hantula, who pioneered the introduction of evolution theory into social science disciplines such as organization science, information systems (IS) research, and economics, wrote: "The most likely scenario is that our brain has been primarily designed to excel in co-located communication, especially where face-to-face interaction takes place" [65, p. iii]. Similarly, Kock argued that "our brain has likely been to a large extent hardwired [i.e., genetically predetermined] for co-located and synchronous communication employing facial expressions" [60, p. 120]. As direct support for these statements, scientific evidence shows that newborns react to human faces in the first hours of life [48]. Clearly, such behavior signifies that the ability to perceive individual faces is innate, and not learned.

The intrinsic human preference toward face-to-face interaction suggests that individuals would not be predisposed to communicate through electronic methods (e.g., e-mail or text-based Internet chat). Evidence based on empirical research supports this view (see, e.g., the works reviewed in [59, 61]). The theoretical arguments underlying this idea, as well as the implications, are summarized under the media naturalness proposition [59], which Kock later refers to as the *media naturalness hypothesis* [60] and the *media naturalness theory* [61]. In essence, this theoretical framework defines the mismatch that exists between the characteristics of face-to-face interaction (which is the benchmark, due to its 100 percent naturalness) and the characteristics of other forms of communication (e.g., e-mail) as independent variables, whereas cognitive effort, communication ambiguity, and physiological arousal are defined as dependent variables.

The degree of naturalness of a communication medium can be evaluated based on the degree to which it incorporates the characteristics of face-to-face interaction [59, 61]. The primary conditions for assessing the level of naturalness include (1) two communicating people share the same context, and they are able to see and hear each other, (2) individuals can quickly (i.e., in real time) exchange communicative stimuli, (3) the situation provides the ability to both convey and observe facial expressions, (4) to convey and observe body language, and (5) to convey and listen to speech. The theory predicts that a decrease in the degree of naturalness (i.e., a higher degree of mismatch) results in three major effects regarding a communications process: (1) increased cognitive effort, (2) increased ambiguity, and (3) decreased physiological arousal (in terms of excitement). Importantly, a low degree of naturalness can negatively affect satisfaction, performance, and productivity in a number of collaborative tasks (e.g., customer support of an online service firm), although in computer-mediated environments humans also have the ability to compensate for lower degrees in naturalness—an ability known as *compensatory adaptation* (e.g., [59]).

To date, both conceptual and empirical articles have investigated the relationship between face-to-face interaction and a number of communications media (see, e.g., research reviewed in [59, 61]). These investigations include e-mail, text-based Internet chat, as well as audio- and videoconferencing. Despite the insights generated by the extensive number of articles investigating these more traditional communications media, our study adds significantly to the extant literature through our focus on the medium of *avatars,* which has gained considerable momentum in use over the past few years, but has not been explored in this regard in the literature.

In human–computer interaction, the term *avatar* is used "as a label for digital representations of humans in online or virtual environments" [6, p. 64]. Typically, avatars are virtual humans that possess realistic faces [29], a fact that is confirmed in several studies published in the IS literature (e.g., [25, 86, 112]). Among the many applications for avatars, both in practical function and as entertainment (e.g., virtual conferences [7], e-learning [34], shopping in virtual malls [73], instant messaging [43], and product innovation [66]), such characters are increasingly used for interactions among humans on the Internet [7, 82]. Against this background, avatars are a ubiquitous phenomenon in today's computerized world, and are important in both private and organizational contexts (e.g., Second Life, World of Warcraft). Because attention to a realistic portrayal of a human face is of particular importance for trust perceptions [117], it has been theorized that the use of avatars may positively affect trust perceptions in computer-mediated interactions, thereby mitigating uncertainty perceptions that are afforded by the anonymity of the Internet. In fact, previous research [12] found that avatar-based communication increases perceived interpersonal trust, relative to the more traditional communications media such as text chat, although an avatar's face does not need to resemble a computer user's real face because it can be chosen freely [87]. Importantly, technologies exist that capture a user's facial expressions via a camera in order to transmit them in real time to the avatar's face (e.g., [120]). Thus, actual facial expressions are transmitted into the virtual world, blurring the boundaries between the real world and cyberspace.

Despite the potential for avatars to induce trust through facial expression, however, evolution theory would suggest that a human face is perceived differently when compared to an avatar face [21, 23, 90, 121], and this is expected to hold true even in a static context (i.e., presentation of pictures of human or avatar faces rather than animations) (e.g., [77, 85, 116, 122]). For human faces, studies indicate that a judgment of reliable trustworthiness can be formed after an exposure time of 100 milliseconds (ms) [121], and because such times are not sufficient for saccadic eye movements, trustworthiness judgments are usually "single glance" impressions [116]. The automatic quality of processing trust judgments suggests that information-processing traits are hardwired (e.g., [59]), which allows the concept of the genetic nature of processing facial information to be extended to trust perceptions.

Against this background, we specifically address three research questions:

> *RQ1: Do differences in trustworthiness discrimination abilities exist, dependent on whether the interaction partner is human or is an avatar?*
>
> *RQ2: If so, are these behavioral differences associated with neural differences?*
>
> *RQ3: Are there differences in learning of trustworthiness in interaction with humans versus avatars?*

The remainder of this paper is structured as follows. In the next section, we present a discussion of the theoretical concepts that are related to our research questions. Based on this discussion, we derive three hypotheses. Specifically, we discuss concepts from evolution theory (H1), brain imaging results from social and cognitive neuroscience (H2), and from brain plasticity theory (H3). Subsequently, we describe the methods that we used to test the hypotheses. An outline of the research results precedes a general discussion of the results. Moreover, we detail our specific contribution to the IS literature. Afterward, we outline limitations of our study as well as possible avenues for future research. Finally, we provide concluding comments.

Theories and Hypotheses

Evolution Theory

Darwin's theory of evolution explains that the human species evolved through natural selection, a process spanning from at least thousands to as much as millions of years, and in which random mutations are introduced in the genetic makeup of offspring, leading to traits that may increase, or may decrease, chances for survival [23, 60]. Those genetic mutations supporting mating and survival are then passed on to offspring, until the mutations become established as species-wide traits. The extremely long neck and legs of the giraffe, for example, are the outcome of such an evolutionary process, because these traits increased foraging abilities in savannah environments (e.g., reaching leaves at the top of trees), thereby increasing the fitness of this species (i.e., the number of surviving offspring). Similarly, perceiving human facial information in order to predict another person's trustworthiness is a trait that has evolved over time, increasing the fitness in environments in which social interaction takes place face to face. Because this trait is generally found in all healthy individuals independent of cultural background, it can be considered as a universal human trait (for details, see Appendix E in [61]).

The universality of the ability to use facial information to predict trustworthiness cannot be automatically extended to avatars. Because these virtual humans have existed for only a few decades, the time period is much too short to lead to behaviorally relevant changes in the human genotype [14, 20, 65]. Drawing on knowledge from evolution theory, the field of *evolutionary psychology* has been developing since the 1980s (e.g., [16, 20, 72]), and has been shown to provide significant “fresh new insights” for IS theorizing [61, p. 395]. In essence, evolutionary psychology posits that a large number of modern brain functions evolved after the emergence of the first hominids, and that these functions, often in an unconscious fashion, significantly affect behavior in the modern world. Relevant contributions to the literature on the potential of evolutionary psychology for IS research include Abraham et al. [1], Kock [62], and Kock and Chatelain-Jardón [64].

The following section briefly explains the functioning of evolutionary psychology theorizing in relation to the focus of this work—the information-processing trait that has evolved to become what we refer to as “face perception.” The explanation is conceptually summarized in Figure 1.

As illustrated at the bottom of Figure 1, genotype influences brain activity. This influence is mediated by the impact of genes on brain structure (e.g., gray matter density), neurotransmitter systems, and other biological mechanisms, because physiology such as brain structure affects both the probability and magnitude of activity in a specific brain area (e.g., [78, 114]). A specific brain activity pattern is, in turn, related to a specific information-processing trait. The present study focuses on the important trait of “face perception.”

From the foundation of evolution theory, it is logical to conclude that an information-processing trait is related to task performance. For example, specific brain modules that direct attention to human faces, and that process facial information, have evolved over time [51, 53], because this has turned out to be advantageous for predicting another individual’s trustworthiness in face-to-face settings. Several parts of the human brain play a crucial role in face perception, such as the pulvinar, inferior occipital gyrus (occipital face area), middle fusiform gyrus (fusiform face area), posterior superior temporal sulcus, amygdala, and anterior infero-temporal cortex (e.g., [18, 51, 53, 104, 119]). Importantly, research on localized brain damage (lesion studies) indicates that damage in the above-mentioned areas is often related to prosopagnosia [15], a disorder in which the ability to recognize faces is impaired, while general cognitive functioning (e.g., decision making) and other aspects of visual processing (e.g., object discrimination) remain intact [104]. This evidence from two streams of research (brain imaging and lesion studies) supports the conclusion that cognitive neuroscience research has developed a substantial knowledge base on the neural foundations of face processing. Moreover, an experiment with both humans and monkeys as subjects shows that primates, in general, “possess the remarkable ability to differentiate faces of group members and to extract relevant information about the individual directly from the face” [22, p. 2973]. This result strongly supports the idea that face perception can be discussed appropriately from an evolutionary perspective, and substantiates the theoretical foundation of the present paper.

The human species, intriguingly, developed a complex system of facial muscles (22 on each side of the face) that allows humans to generate more than 6,000 expressions for communicating thoughts and feelings [8, 58, p. 10, 60, p. 120, 81]. Trustworthi-

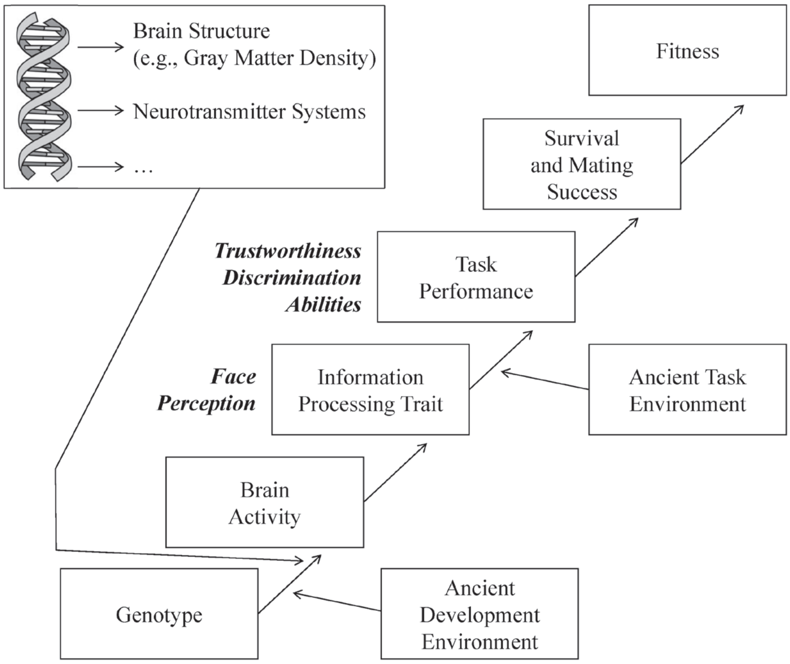


Figure 1. Evolution of Information Processing Trait by Natural Selection

Source: Based on Kock [61, p. 398].

Notes: Complementary to the original theorizing, we added the element "brain activity," thereby signifying that the influence of genotype on a specific information processing trait is mediated by activity in specific brain areas (e.g., [50]). Moreover, we added another element illustrating that the influence of genotype on brain activity is mediated by brain structure (e.g., gray matter density), neurotransmitter systems (i.e., chemicals that transmit signals from a neuron to another one across a synapse), and other neurobiological mechanisms (e.g., [78, 114]). Also, we have added both the information processing trait "face perception," and the corresponding ancient task performance, namely trustworthiness discrimination abilities. Altogether, this figure conceptually integrates the brain, as well as the specific topic of this article—face trustworthiness—into evolutionary psychology theorizing.

ness discrimination abilities, based on encoded facial information, are thought to have increased survival success and mating success, increasing the likelihood that the behavior trait would be replicated in the succeeding generation, with a positive effect on fitness as a result.

Moreover, Figure 1 demonstrates that the influence of genotype on the brain has been affected by the ancient development environment in its broadest sense, ranging from nutrition to social conditions [61]. In general, there is agreement among scientists that most traits that are significantly affected by genetic makeup, including those related to face-to-face communication, need interaction with the environment to become completely developed (e.g., [17, 59]), a phenomenon referred to as *epigenetic.* With respect to the ability to perceive individual human faces, for example, early visual input (e.g., during the first few months of life) is a necessary precondition for the normal development of neural mechanisms that will later specialize to support the processing of faces [74].

Furthermore, as is illustrated in Figure 1, the ancient task environment moderates the influence of an information-processing trait on task performance [61]. Because the ancient task environment was characterized by the nonexistence of technology for communication during almost 100 percent of human history (written communication emerged as recently as approximately 5,000 years ago, and electronic forms of communication were not invented until the very recent past), attention to human faces and understanding facial features has been indispensable for effective communication and for survival until the recent past. Understandably, millions of years of exercise in face-to-face interaction would have resulted in elevated abilities to use facial information to discriminate between trustworthy and untrustworthy actors. In contrast, if humans had interacted only by way of technologies such as telephone or text chat that suppress facial information (task environment), perceptive reading of the human face (information processing trait) would not likely have developed significant trustworthiness discrimination abilities (task performance).

Against the background discussion herein, and through an application of evolution theory and evolutionary psychology, we derive the following hypothesis:

> *Hypothesis 1: Based on facial information, people are better able to predict the trustworthiness of humans than the trustworthiness of avatars.*

H1 pertains to the behavioral level of analysis. However, because all human behavior is—at least partly—determined by biological factors, in particular those related to genes and the brain (e.g., [17]), there is reason to assume that behavioral differences are accompanied by neurobiological differences.

Trust, Mentalizing, and the Medial Frontal Cortex

The concept of *NeuroIS,* recently introduced in the IS literature, describes the idea of applying neuroscience theories, methods, and tools in IS research [27]. Among other contributions (e.g., [27, 28, 97, 100, 102, 118]), this new research field is expected to add to the development of new theories that make possible accurate predictions of behavioral responses to information technology (IT) [102]. Such new theories seek to integrate the behavioral and biological levels of analysis, thereby leading to a deeper understanding of IS phenomena. Specifically, functional magnetic resonance imaging (fMRI) studies have demonstrated that brain activation patterns have the power to explain differences in behavior toward IT artifacts (e.g., [11, 26, 99]).

In this context, we would expect that the behavioral difference specified in H1 is accompanied by activation differences in specific brain areas. Parts of the *medial frontal cortex* (MFC), in particular, have been identified as crucial brain regions for the neural implementation of trustworthiness predictions (e.g., [71, 79]). Specific brain areas within the MFC, which correspond approximately to Brodmann areas (BAs) 9, 10, 24, and 32, facilitate neural implementation of mentalizing, a human ability to infer the internal states of other actors (e.g., intentions) in order to predict their personality traits and behavior (see, e.g., the following reviews on the importance of the MFC for mentalizing: [2, 4, 40, 42, 44]).

In the scientific literature, the concept of mentalizing is also referred to as the *theory-of-mind* [41, 91, 109]. In general, mentalizing is a fundamental cognitive process in trust situations because the decision to trust involves thinking about an interaction partner's intentions to infer his or her trustworthiness. Fehr confirms this view: "Since trust decisions are also likely to involve perspective-taking, they should also activate areas implicated in theory-of-mind tasks" [37, p. 228]. A recent review of the literature on the biology of trust [98] identifies the concept of mentalizing and the corresponding MFC activations as significant determinants of human trust behavior. Importantly, this review integrates research from various scientific fields such as neuroeconomics, social neuroscience, and NeuroIS, which supports the theory that mentalizing and MFC activations are fundamental in trust situations across many different contexts.

Based on results of a brain imaging experiment in which participants had to judge *human* faces regarding their trustworthiness, Winston et al. conclude that "social judgments about faces reflect a combination of brain responses that are stimulus driven . . . and driven by processes relating to inferences concerning the intentionality of others" [122, p. 281]. Similarly, Frith and Frith argue that "faces, in particular, are an important source of information about their inner states. For example, there is agreement about what a trustworthy person looks like" [40, p. 531]. Finally, a review on the neurobiological foundations of mentalizing indicates that "[i]n the case of knowledge of other minds, we appear to begin with . . . perception of a face" [3, p. 696]. These studies suggest that the human face provides rich information for predicting another individual's intentions and behavior. Considering that face-to-face interaction has been the major communications mode for the majority of human history, a specific trait for processing human facial information and corresponding trustworthiness prediction abilities must be part of the human genetic makeup. A number of other scholars share this view; they also regard trust and mentalizing as "evolved capacities" (e.g., [38, p. 277, 115, p. 201]). With respect to possible explanations of *why* such perception of human faces constitutes a basic instinct—and as a complement to the fact that trust in untrustworthy humans could have directly resulted in death—scholars have offered explanations related to the effectiveness of social functioning (for details, see [32, 95, p. 65]).

Building on the brain imaging research presented in this section, and extended by the knowledge that brain activation mediates the influence of genes on behavior (e.g., [17, 50]), we derive the following hypothesis:

> *Hypothesis 2: When determining whether or not to trust another actor, decision making activates the MFC more during interaction with humans, if compared to interaction with avatars.*

Brain Plasticity Theory

So far, we have argued that because the ability to accurately assess another person's trustworthiness on the basis of facial information has increased survival probability over the course of millennia, the human brain has been shaped by that evolution process. However, as explained by scholars in various disciplines such as biology, psychology, and neuroscience (for a review of sources, see [59, p. 336]), evolution not only has given the human species a genetic makeup with a predetermined set of instincts (such as trustworthiness discrimination abilities based on face perception), but has also equipped humans with a "plastic" brain. Plasticity of the brain refers to its ability to change structure and function [67]. Obviously, this general characteristic of the brain is advantageous for survival, thereby increasing fitness. Learning and memory, as well as experience, are concepts that are inseparably associated with brain plasticity, because they imply changes in dendritic length, synapse formation, and metabolic activity, among other modifications [69]. Such neurobiological changes often lead to changes in behavior [67].

The ability of the human species to develop new schemas in the brain pertains to most types of schemas, including those related to computer-mediated communication [59]. Against this background, it is no surprise that predictions of behavioral theories about electronic forms of communication are consistent with predictions based on brain plasticity theory. Kock, for example, explains that channel expansion theory [19] "is compatible with brain plasticity" [59, p. 336]. This theory is in line with the prediction that continued use of an unnatural communication medium (e.g., avatar-based interaction) may, over time, result in learning processes. Such processes may, importantly, lead to lower levels of communication ambiguity, a characteristic that is usually associated with face-to-face communication but is not typically associated with interaction without real facial information.

In this context, Kock coined the phrase "cognitive adaptation proposition" [59, p. 336]. To illustrate the concept (in a later study), Kock cited Carlson and Zmud [19, p. 157] on the example of e-mail: "As individuals develop experience communicating with others using a specific channel, such as e-mail, they may develop a knowledge base for more adroitly applying this communication channel. . . . [U]sers may become aware of how to craft messages to convey differing levels of formality or of how to use channel-specific metalanguage to communicate subtleties. Similarly, these individuals are also likely to interpret messages received on this channel more richly because they can interpret an increasing variety of cues" [61, p. 410]. This statement supports a clear understanding that schemas that are not in the human genetic makeup, but can be learned through direct experience, may affect behavior to a considerable degree. Kock and Hantula even suggest that, due to brain plasticity, electronic forms of communication may become "second nature" [65, p. iii]. Similarly, Schlicht et al., investigating trustworthiness perceptions of opponents' faces while playing poker, argue for the importance of learning mechanisms: "Once rapid impressions have been formed, beliefs can later be updated by direct experience with the individual, to develop a new estimate that will be used going forward" [106, p. 1].

In general, scholars (e.g., [68, pp. 718–746]) indicate that learning is always accompanied by changes in the nervous system. Consequently, learning can be studied at various analytical levels, ranging from observable behavioral changes to alterations in molecular structure. Applying the context of brain plasticity theory, we derive the following hypothesis, which pertains to the behavioral level of analysis:

Hypothesis 3: In a trust situation, people's learning rate of trustworthiness is similar, whether interacting with humans or avatars.

Methods

Tasks and Measurement

In order to test our hypotheses, we used a modified version of the original trust game [13], a multiround trust game in which an investor has an initial endowment. The investor first decides whether to keep her or his endowment, or to send it to the trustee. Then the trustee observes the investor's action (i.e., sending or not sending) and decides whether to keep the amount received or to share it with the investor. The experimenter multiplies the investor's transfer by some factor, so that both players are advantaged, collectively, if the investor transfers money and the trustee sends back a part of it. As a behavioral measure for trust we used the decision of the investor to send money, and as a behavioral measure for trustworthiness we used the trustee's decision of whether or not to return money [36].

The trust game has been used as an experimental paradigm in a number of investigations, particularly in studies on the neurobiological foundations of trust (for a review, see [98]). Moreover, this game has recently been described as an appropriate task for the study of mentalizing in trust situations within brain imaging environments. Sripada et al. wrote: "The medial prefrontal cortex has been implicated as a key brain region that implements mentalizing during the social interactions [and] the 'trust game' serves as a potent probe of mentalizing abilities because it sets up the need to make inferences about the mental state of others. . . . Therefore, functional neuroimaging of the trust game . . . may shed light on a novel behavioral and neural mechanism" [111, p. 984]. Thus, the trust game, if used in an fMRI environment, is a suitable task for investigating both H1 and H2.

In a research agenda addressing trust in online environments, Gefen et al. wrote that "[t]rust develops gradually as people interact with each other. . . . [T]herefore it is also important to study the longitudinal effects of trust on transaction decisions and other behavioral outcomes" [46, p. 277]. The trust game can be played either on a one-shot basis or as a multiround version. In the latter case, two actors play against each other multiple times (e.g., ten consecutive rounds; [55]). As a result of this extended interaction, the longitudinal nature of trust is accentuated and players can learn about the behavior of the other actor. We used such a multiround version of the game, as this is a precondition for the investigation of H3.

Stimulus Material

In our trust game experiment, the participants played the role of investor against both humans and avatars in the role of trustee. The objective of the stimulus selection was therefore to identify a number of human and avatar faces for the fMRI study. Ultimately, both the human and avatar groups comprised four faces with a high degree of trustworthiness and four faces with a low degree of trustworthiness (examples are provided in

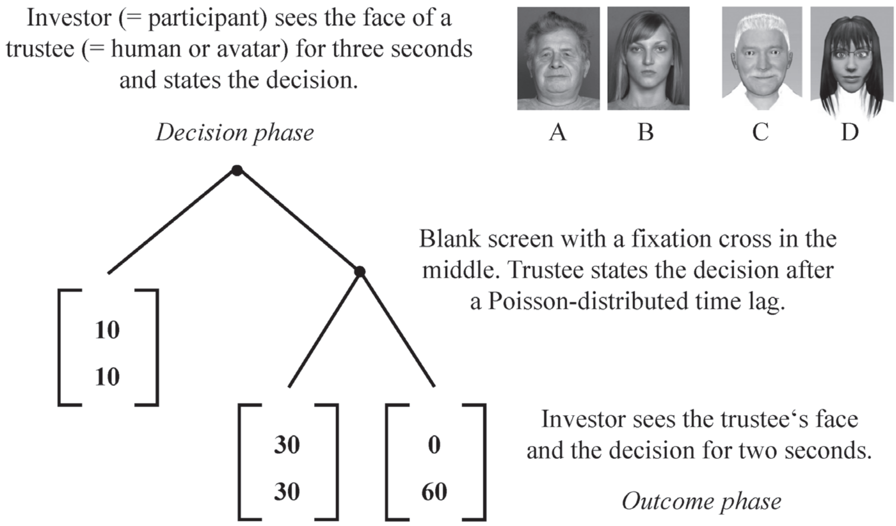


Figure 2. Structure and Payoff Matrix of the Trust Game

Notes: The upper value in the square brackets indicates the investor's payoff, and the lower value shows the trustee's payoff. Participants played against eight humans and eight avatars in the fMRI study. The human and avatar groups were matched for pre-rated trustworthiness based on facial appearance, and for behavioral trustworthiness (i.e., the number of rounds in the trust game in which money was returned). We used four trustworthy humans (see, e.g., Face A) and four untrustworthy humans (e.g., Face B), as well as four trustworthy avatars (e.g., Face C) and four untrustworthy avatars (e.g., Face D). All humans and avatars had a direct gaze, as illustrated in the four example faces. Both the human and avatar groups consisted of four males (two trustworthy and untrustworthy ones) and four females (two trustworthy and untrustworthy ones). The complete set of 16 faces used in our fMRI study is available elsewhere [101].

Figure 2). The complete set of 16 faces used in our fMRI study is available in Riedl et al. [101], as is detailed information on the stimulus material selection process.

Subjects

For the main study, we selected 11 male and 8 female subjects (none of whom had participated in the stimulus selection pretest). All of the subjects were healthy and reported no history of neurological or psychiatric diseases. The mean age of the subjects was 31.83 years. One subject (female) had to be excluded from further analyses, as she indicated after the fMRI session that she was afraid while in the fMRI machine, and as a result was not really thinking of the assigned task. Therefore, the sample size underlying our data analysis is $N = 18$. All of the participants were familiar with the Internet and had been using it for many years (mean = 10.77, SD [standard deviation] = 3.40, minimum = 4, maximum = 18). By design, our investigation is focused on experienced computer and Internet users rather than novices. All of the participants were paid for their participation, and gave written informed consent. The study was approved by the Freiburg Ethics Commission International (FECI), Germany. Further details on the subjects' characteristics are described elsewhere [101].

Experimental Procedure and Stimulus Presentation

In our experiment, the participants played the trust game in the role of the investor against (1) humans and (2) avatars (both playing in the role of trustee, see Figure 2). Our game, therefore, mimics a typical interaction in both social and economic exchange relationships, both in bricks-and-mortar environments (human condition) and in computer and online environments (avatar condition). The participants were told in advance that their playing partners (i.e., the trustees) would not be responsive to their playing strategies. We did stress, however, that each trustee has a specific character that determines his or her trustworthiness. Half the trustees, both humans and avatars, were predetermined by the experimenters to be relatively trustworthy, whereas the other half were predetermined as relatively untrustworthy. Trustworthy actors returned €30 in seven out of ten rounds, whereas untrustworthy actors returned €30 in only three rounds. Apart from their facial appearance, the participants had no information regarding the trustees.

In each round of the game, the participants were asked whether they wanted to keep their initial endowment of €10 (≈ $13 at the time the experiment was conducted), or whether they wanted to give it to the trustee whose face was presented to them. In the case of giving the €10 to the trustee, the amount was multiplied by six (resulting in €60), which the trustee could then either keep, or split (i.e., return €30 to the participant). The participants played ten rounds of the game with each trustee, with three seconds in each round to make the investment decision (= *Decision phase;* see Figure 2); note that decision times of one to three seconds are sufficient to make economic decisions in an fMRI environment (e.g., [122]). After a variable time (varied based on a Poisson distribution) in which a blank screen with a fixation cross in the middle was presented, the trustee's face and the decision was visually presented for two seconds (= *Outcome phase;* see Figure 2). Before the first round, after the fifth round, and after the tenth round, the participants were asked to rate the trustworthiness of the trustee, which was operationalized as the probability that the trustee, in the case of being given money, would behave in a trustworthy manner (i.e., returning €30).

We used the program Presentation (Neurobehavioral Systems, Albany, CA) to present the stimuli in the MRI-scanner (3T Siemens Tim Trio, Erlangen, Germany) and to record the responses on a laptop computer with Windows XP as the operating system. Visual stimuli were presented using video goggles. Details about data collection and analysis are provided in the Appendix.

Results

Before we tested H1, and without making a distinction between trustworthy and untrustworthy actors, we analyzed whether or not the participants differ in their decisions to trust (i.e., to invest their initial endowment of €10) when playing against humans or avatars. In essence, we found that the participants showed considerable trust behavior, independent of playing against humans or avatars (average trust in the human condition: in 52 of 80 games, SD = 11.55; average trust in the avatar condition: in 51.39 of 80 games, SD = 13.57). There was no significant difference between the number of decisions to trust in the human and avatar conditions ($t = 0.445$, df [degrees of freedom] = 17, $p = 0.662$).

Hypothesis 1

In both the human and avatar groups, discrimination between trustworthy and untrustworthy actors revealed a significant difference in trustworthiness prediction ability, as was predicted in H1. We used the trustworthiness evaluations of the pretest (see [101]) as a benchmark (an established procedure in neuroscience face evaluation studies [105, p. 2]). Similar to the results of the pretest, we found a significant ($p < 0.1$) difference between high and low trust faces in the fMRI trustworthiness ratings, indicating a positive manipulation check.

While we found no significant difference in the participants' decisions to trust when playing against trustworthy versus untrustworthy avatars (25.89 versus 25.50, $t = 0.408$, df = 17, $p = 0.689$), we identified a significant difference when playing against trustworthy versus untrustworthy humans (27.61 versus 24.39, $t = 3.502$, df = 17, $p = 0.003$). The difference between the difference values of the number of decisions to trust humans, trustworthy versus untrustworthy (Δ 3.23), and to trust avatars, trustworthy versus untrustworthy (Δ 0.38), is statistically significant ($t = 2.536$, df = 17, $p = 0.021$). Thus, the participants trusted *un*trustworthy avatars to a similar degree as they trusted trustworthy avatars. In the human group, importantly, we could not observe this disadvantageous behavior. Our results therefore indicate that people are better able to predict the trustworthiness of humans than the trustworthiness of avatars. This finding confirms H1.

Hypothesis 2

In the present experiment, we focus on the investigation of brain activation during the *Decision phase* (see Figure 2); that is, we analyzed brain activation during the presentation of the trustees' faces. In this phase, the participants had to decide whether or not to send their initial endowment of €10 to the trustee.

We found that activation in the dorsomedial prefrontal cortex (DMPFC), the rostral part of the anterior cingulate cortex (rACC), and the ventromedial prefrontal cortex (VMPFC) was significantly higher in the human condition compared to the avatar condition ($z > 3.09$; cluster size > 25; see Figure 3 and Table 1). These three brain regions correspond approximately to BAs 9, 10, 24, and 32, and they have been summarized under the label "MFC," a brain structure that plays a significant role in the neural implementation of mentalizing [2, 4, 40, 42, 44]. This finding confirms H2.

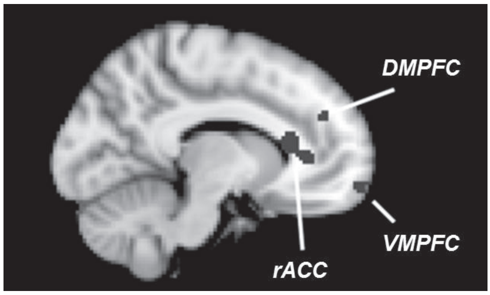


Figure 3. Brain Activation Results (Sagittal Cut)

Notes: Brain activation in the *Decision phase* (fMRI contrast: human–avatar). The reverse contrast (avatar–human) did not result in significant activation differences. The three brain areas play a significant role in the neural implementation of mentalizing. DMPFC = dorsomedial prefrontal cortex, rACC = rostral anterior cingulate cortex, VMPFC = ventromedial prefrontal cortex.

Hypothesis 3

In order to understand the participants' trust learning process, particularly their learning rates, we modeled the perceived trustworthiness of the trustees on a round-by-round basis. The underlying assumption of this approach is that perceived trustworthiness of a trustee is based on the facial appearance before the first round begins, and is adjusted based on its behavioral trustworthiness during the subsequent rounds of the game. Importantly, recent empirical evidence provides support for this assumption [105].

We assume that perceived trustworthiness is reflected in the subjective probability that a trustee will be trustworthy (i.e., will return €30) when a participant invests the initial endowment of €10. Moreover, a participant typically selects the gaming strategy that offers a higher expected value. The expected value of keeping the initial endowment of €10 is €10, because the probability for this gain is 100 percent in our game (see Figure 2, left path). The expected value of investing the endowment of €10 is equal to the product of a trustee's perceived trustworthiness and the possible payoff of €30. Hence, the trust decision depends only on the perceived trustworthiness (see Figure 2, right path).

During the ten rounds of interaction with a specific trustee, trustworthiness is expected to be updated on the basis of that trustee's behavior. In this paper, we model the updating process with a *reinforcement learning model* [9, 10, 96, 113]. Such a model generally assumes that after the decision has been made for one alternative, a received reward $R(t)$ at time t is compared to an expected value $EV(t)$, with the deviation d formalized as prediction error PE: $d(t) = R(t) - EV(t)$. A reinforcement learning model assumes that learning is driven by these deviations; hence, a PE is used to update $EV(t)$, allowing the optimization of reward predictions. The influence of a specific PE on $EV(t)$ regarding the next trust decision is determined by the learning rate.

Table 1. Brain Areas Activated in the Decision Phase (Human–Avatar)

Anatomical region	Cluster size (voxels)	Maximum z-score	MNI coordinates (x, y, z)		
Rostral anterior cingulate cortex	**163**	**4.23**	**–8**	**26**	**14**
Lateral occipital cortex	56	3.70	–38	–66	24
Precuneus/posterior cingulate cortex	49	3.97	12	–54	44
Temporal pole/lateral orbitofrontal cortex	37	3.79	30	20	–28
Precuneus/posterior cingulate cortex	36	3.65	–16	–54	38
Frontal pole/ventromedial prefrontal cortex	**32**	**3.90**	**–8**	**66**	**–12**
Dorsomedial prefrontal cortex	**28**	**3.67**	**–8**	**44**	**28**
Frontal pole	28	3.56	–8	42	–28

Notes: MNI = Montreal Neurological Institute. There was no significant activation in the reverse contrast (avatar–human). The brain regions indicated in bold are summarized under the label medial frontal cortex (MFC); they are discussed in this article because they play a significant role in the neural implementation of mentalizing, a concept of high relevance in trust situations (see, e.g., a review by Riedl and Javor [98]).

Formally, a reinforcement learning model is defined as

$$EV(t) = EV(t-1) + \alpha \cdot d(t-1). \tag{1}$$

From $EV(t)$, it is possible to calculate the subjective probability of trustworthy behavior at time t by dividing $EV(t)$ by €30. In our trust game, a model with a constant learning rate (see formula (1)) would assume that the perceived trustworthiness is updated equally for trustworthy and untrustworthy behavior.

We fitted the reinforcement learning model for the interaction with humans and avatars, and instructed each participant to rate the trustworthiness of each trustee before the first round of the trust game began (based solely on facial appearance). We used these initial ratings as the starting points of the reinforcement learning process. The free parameter α was fitted by minimizing the sum of squared differences between model predictions and a participant's trustworthiness ratings after the fifth and tenth rounds.

We investigated whether learning rates were significantly different between the human and avatar conditions (see Table 2 for a summary of the statistics). As expected, we found no significant differences ($t = 0.1510$, df = 17, $p = 0.8820$), thus confirming H3.

Assuming a deterministic decision strategy specifying that people always choose the alternative that offers the higher expected value when two alternatives offer a higher-than-expected value, the reinforcement learning model, on average, correctly predicts 73.61 percent (71.51 percent) of the trust decisions in the human (avatar)

Table 2. Trustworthiness Learning Rates (Human and Avatar)

Trustworthiness learning rates	Mean (SD)
Sum of squared differences (constant learning rate, human)	0.2343 (0.1322)
Sum of squared differences (constant learning rate, avatar)	0.2773 (0.1950)
Alpha (human)	0.2139 (0.0825)
Alpha (avatar)	0.2122 (0.0916)
Correct model predictions (constant learning rate, human)	73.61% (9.40%)
Correct model predictions (constant learning rate, avatar)	71.51% (13.92%)

condition. Importantly, the predictive power of this model is significantly higher than chance level (human: $t = 10.6540$, df = 17, $p < 0.001$; avatar: $t = 6.870$, df = 17, $p < 0.001$). Further analysis regarding the robustness of our results related to H3 is provided elsewhere [101].

General Discussion

Measuring by means of fMRI in a multiround trust game, we investigated (1) individuals' trustworthiness discrimination abilities in their interaction with humans versus avatars (H1); (2) the underlying neurobiological mechanisms, with a focus on activation in the MFC, a brain area well known for its importance in the neural implementation of mentalizing (H2); and (3) individuals' trustworthiness learning rates (H3). First, based on theories from evolutionary psychology, we hypothesized that people are better able to predict the trustworthiness of humans than the trustworthiness of avatars. Second, based on findings from brain imaging studies in social and cognitive neuroscience, we predicted that making decisions about whether or not to trust another actor will activate parts of the MFC significantly more during interaction with humans, if compared to interaction with avatars (in this context, also see related research from the field of robotics [56, 70]). Third, based on brain plasticity theory, we predicted that, whether interacting with humans or avatars, the trustworthiness learning rate of people is similar. We found support for all three hypotheses.

Trust evaluations that are based on the processing of facial information, as well as corresponding brain activity patterns, are phenomena in social interactions that provided a survival advantage during the past eras of human history. Although only recently have computer-mediated forms of communication and cooperation begun to emerge, they promise to bring revolutionary changes in the organization and interaction of societies. Of particular interest for our research, the exponential increase in the number of individuals who communicate and cooperate via computers and the Internet, and who are beginning to represent themselves in virtual worlds through avatars, signifies a changed understanding of interaction [7]. Genotype, which affects the development of social skills such as mentalizing and trustworthiness predictions, evolved in natural, face-to-face environments. Today's computerized world, however, often lacks natural facial information. Rather, people are afforded the possibility of representing themselves as avatars, thereby hiding their real faces behind a mask, which presents opportunities to actively misrepresent themselves [43]. Therefore, although interaction on the Internet may have benefits (e.g., [54]), the lack of real human faces in communication may serve to reduce these benefits, in turn leading to reduced levels of collaboration effectiveness. A main reason for this resulting condition is that the human brain is preprogrammed for face-to-face interaction, but not for computer-mediated interactions—in the words of evolutionary psychologist David M. Buss, humans "carry around a stone-aged brain in a modern environment" [16, p. 20].

However, despite the fact that some hardwired instincts have made humans less adaptive to life in a computerized world [65], one remarkable feature of the human brain is its flexibility. Interestingly, this powerful feature is itself an outcome of evolution [60], and this flexibility has been described as "mental meta-module . . . favored by natural selection" [63, p. 26]. In situations in which the environment is changing too fast for genetically coded instincts to be patterned in the brain, humans can nevertheless adapt to a situation in short time periods, because experience constantly changes the structure and/or functioning of the brain, leading to behavioral adaptions (e.g., [110]). Importantly, a change in the brain developed through experience is not a phenomenon that exclusively pertains to children; rather, the adult brain is also "plastic" (e.g., [67, 69]).

Despite this cognitive adaptation ability, however, the major differences between hardwired instincts and learned schemas is that the former require less cognitive effort (brain activity), and typically operate on an unconscious level (e.g., [75, 105]). This, in turn, leads to more freely available cognitive (brain) resources that may be used for other purposes, or, if not used, people may experience a lower level of cognitive load. Both factors, obviously, may increase efficiency. However, even though learning may compensate for deficits in hardwired abilities, this compensation does not come without a price (e.g., cognitive effort). Kock, citing works from evolutionary psychology, confirms this view: "As far as human communication is concerned, learned circuits are unlikely to be as efficient as the hardwired circuits endowed on us by evolution. . . . [G]enetically coded circuits are . . . automatic, unconscious, and undistracted by irrelevant aspects of world knowledge" [60, p. 121].

Specific Contributions to the IS Literature

In addition to the contributions discussed in the previous section, which refer to people's interaction with humans and avatars in general, and are therefore related to the scientific discourse in various disciplines (e.g., human–computer interaction, psychology, cognitive neuroscience), the present paper makes four specific contributions to the IS literature. First, we integrate two previously unconnected IS research domains based on the example of avatar research, namely IS evolutionary psychology and NeuroIS. Second, we generate novel insights within the domain of IS avatar research. Third, we create a knowledge base that is relevant for IS trust research. Fourth, in addition to these three theoretical contributions, this paper makes a methodological contribution to the IS literature. We discuss these four contributions in the following subsections.

Contribution 1

In this paper, we connect two previously unconnected IS research domains, namely, IS evolutionary psychology [61] and NeuroIS [27, 28, 97, 100, 102, 118]. Evolutionary psychology, in general, argues that a large number of modern brain functions evolved during the past millions of years, and that these functions, usually without conscious perception, significantly affect human performance and behavior in the modern, highly computerized world. Recently, IS researchers tellingly argued: "[l]et us remember that the person to whom we give today's latest IS has an operating system that has changed little in [the past] 100,000 years [and therefore] humans are the ultimate legacy system" [1, p. 68]. Arguments such as this substantiate the value of evolutionary psychology for IS theorizing.

Importantly, influential theoretical articles in the IS literature (e.g., [61]) as well as empirical contributions (e.g., [64]) have been published on the subject of evolutionary psychology, but to the best of our knowledge no previous IS study has applied brain imaging technology, based on a behavioral research paradigm (the trust game), to directly investigate the neurobiological *and* behavioral consequences of people "carry[ing] around a stone-aged brain in a modern environment" [16, p. 20]. Unlike previous evolutionary psychology research in the IS domain (for a collection of corresponding articles, see [62]; on the value of evolutionary psychology theorizing in the context of technology acceptance research, see [1]) that has *not* investigated brain activity as a mediating factor between genotype and a specific information-processing trait (here perception of human faces in the context of avatar research), we used fMRI to shed light on the nomological network underlying evolutionary psychology theorizing (see Figure 1). This contribution, importantly, is in line with the current call for the application of neuroscience theories and tools to inform IS phenomena (e.g., [27, 28, 102]), and specifically, our study demonstrates how the behavioral *and* biological levels of analysis can be integrated to better understand the antecedents of human behavior toward IT artifacts such as avatars.

Generally, we investigated two constructs that have been identified as major variables in IS research—trust and mentalizing [27, p. 691]. Our comparative study (human versus avatar) therefore contributes to a better understanding of two major IS concepts. Moreover, our contribution is in line with a previous review that identified a "neuropsychological focus" and an "evolutionary psychology lens" as important for IS avatar research to progress [107, pp. 440, 443].

Contribution 2

Based on an extensive analysis of the IS literature, we could identify only a limited number of avatar papers (the analysis may be obtained in electronic form by request from the first author). Thus, avatar research, in general, is an underrepresented topic in the IS literature, and consequently needs more attention within the IS discipline. Against such a research background, our study addresses a topic that has hardly been explored, yet is a highly timely topic that has become increasingly important during the past years, mainly as a consequence of the rising pervasiveness of virtual worlds [108].

Moreover, the present paper generates valuable insights within the domain of contemporary IS avatar research. Our analysis of the IS avatar literature reveals that most empirical investigations manipulated avatar design features (independent variables) in order to study the resulting effects on important outcome variables. For example, Qiu and Benbasat [93] manipulated humanoid embodiment (avatar versus no avatar) and output modality (human voice versus text-to-speech versus text) in order to study the effects on usage intention, a relationship that is mediated by social presence and trusting beliefs. As another example, Hess et al. [52] studied the effects of an avatar's vividness (among other features) on trusting beliefs, and hypothesized that this relationship is mediated by social presence. To state a third example, Nunamaker et al. [89] investigated how an avatar's gender and demeanor affects perceptions of the avatar's power, trustworthiness, likability, and expertise. In addition to these experimental studies, qualitative research has been carried out. Mueller et al. [86], for example, interviewed employees of IBM (who use the virtual world Second Life as a corporate knowledge management platform) in order to identify differences between interaction based on more traditional technologies (e.g., e-mail) and interaction via avatars regarding "knowledge and knowing activities." Thus, this kind of research compares human interaction via avatars to other forms of electronic interaction.

Overall, we conclude from our literature review that previous IS avatar research either studied the effects of design manipulations on outcome variables or compared interaction via avatars with other forms of electronic communication. Despite the valuable insights that these studies have generated, a much more fundamental question has received little analysis in the IS literature, namely, the question about the differences between human interaction in face-to-face settings and human interaction via avatars (we could identify only one study [39], discussed below). We believe that examination of this question is fundamental, as IS research should not only compare the effects of the different forms of electronic communication (e.g., e-mail, text chat, videoconferencing, avatar) but should also compare these various forms of electronic communication (e.g., avatar) with people's natural interaction method, namely, face-to-face communication.

The significance of this topic, importantly, is even reflected in the IS avatar literature itself, as signified by the following statements. Mueller et al., for example, wrote: "[Second Life] allows to share an interactive synthetic environment from a true first-person perspective, which is regarded as the second best alternative to the direct interaction of real people" [86, p. 490]. As another example, Nunamaker et al. indicated that "[w]hile humans are relatively good at identifying expressed emotions from other humans whether static or dynamic, identifying emotions from synthetic faces is more problematic" [89, p. 22]. Similarly, Franceschi et al. argue that "[e]ffective group collaboration requires . . . the existence of an appropriate working environment—an environment that makes all group members feel comfortable and fosters their participation. This is more difficult in distance group collaboration due to the lack of face-to-face contact . . . facial expressions provide rich information about the emotions and intentions of others. For this reason, probably, executives continue to travel thousands of miles to have face-to-face meetings" [39, p. 84]. Finally, as further confirmation of the value of the present study, Davis et al. explain that "[w]hether personal focus or direct contact is the same in a metaverse as in face-to-face environments is yet unanswered" [25, p. 96] (a metaverse is a three-dimensional virtual world in which people interact via avatars). Considering these statements, as well as the fact that human interaction in face-to-face settings has hardly been compared to human interaction via avatars, our study contributes to closing a significant research gap in the IS avatar literature.

As mentioned, we could identify only one article in which face-to-face interaction was directly compared to human interaction via avatars [39]. Specifically, in this field study, three learning environments (a traditional face-to-face classroom setting, a text-based virtual learning environment such as Moodle, and an avatar interaction in a virtual world based on Second Life) were compared regarding their effects on student performance. The major result of the study is that "performance of participants in the virtual world learning environment is significantly better than that of the other two environments" [39, p. 94]. This result is not in line with our finding regarding H1, where we discovered a performance advantage for people's interaction with humans, when compared to interaction with avatars (regarding trustworthiness discrimination abilities). However, the result indicating that students perform better in an avatar environment than in a face-to-face setting might be an artifact of sample selection; Franceschi et al. [39] recognize that a "novelty bias" might have affected their results. Specifically, "we mean that a certain technology, such as virtual worlds, attracts attention simply because it is new. This could potentially affect the students' measured engagement [with potential subsequent effects on performance] . . . admittedly, most of the students in our study were technology novices" [39, p. 96]. However, it might also be possible that their result reflects the brain's flexibility to quickly adapt to new environments, thereby supporting our finding regarding H3. Future studies based on more sophisticated experimental designs are therefore necessary for developing more definitive conclusions. Overall, however, the results of our study suggest that the use of avatars for human interaction and collaboration (e.g., virtual learning) may be less beneficial than previously assumed—a fact that is of particular relevance for managers, software engineers, and policymakers, among others.

Contribution 3

The present paper generated knowledge relevant for IS trust research. Specifically, our study responds to a call in a prominent research agenda paper for more investigation into the "longitudinal nature of trust" [46, p. 277]. In this research agenda for trust in online environments, Gefen et al. explain that "[t]rust, of course, is not only about one-time interactions. Trust develops gradually as people interact with each other. . . . [T]he importance of trust as a key consideration decreases with experience [and] models should have examined how the importance of trust changes over time" [46, p. 277]. Importantly, Qiu and Benbasat [93, p. 166] made a similar recommendation for additional studies employing longitudinal designs, and their call is directly concentrated on avatar trust research. Obviously, such calls confirm the value of the present paper, particularly with respect to H3.

In this context, it is also important to discuss both swift trust and knowledge-based trust, two concepts that were studied in face-to-face settings and virtual teams [103]. Swift trust denotes high levels of initial trust in another individual without any knowledge about that person, whereas knowledge-based trust denotes trust developed through interactions, based on the assessment of actual behavior. In their study, Robert et al. integrated these "two seemingly contradictory views of trust" [103, p. 242]; specifically, they manipulated team member characteristics as well as team member behavior to empirically test a theoretical model of trust formation and the influence of IT on trust formation. One major finding of their study was that perception of team member characteristics (ability, integrity, benevolence) dominated the initial formation of swift trust. However, once individuals gained knowledge, swift trust became less important and knowledge-based trust became dominant.

The results of the present paper are in line with the findings reported in Robert et al.'s [103] study; that is, subjects learned the trustees' trustworthiness based on repeated interaction (knowledge-based trust), and this affects the relevance of initial trustworthiness assessments (swift trust). A major difference between these two examinations is, however, the methodological approach to study swift trust and knowledge-based trust. While Robert et al. used "simulations of real events" (referred to as "vignettes" in the scientific literature [103, p. 251]), we investigated *actual behavior,* based on the trust game. Thus, our study overcomes a major limitation of Robert et al.'s study: "subjects could potentially respond differently to a hypothetical scenario [and] vignettes are sometimes not as powerful a manipulation as traditional experiential experiments" [103, p. 267]. However, despite the fact that vignettes cannot substitute for the investigation of actual behavior, Robert et al.'s study makes a substantial contribution to the IS literature, and we consider our experiment as a valuable complement to this prior investigation. Moreover, the value of our study is further substantiated by Robert et al., who indicate in a section on the implications of their study for future research that "we believe that more research is needed on swift trust and the *factors* that influence initial trust judgments before knowledge of behaviors has been gained" [103, p. 267, emphasis added]. Obviously, an interaction partner's face—the thematic focus of the present paper—is such a factor.

Contribution 4

Neurobiological research on trust is typically based on definitions and measurements that are different from the conceptualizations used in the IS literature. McKnight and

Chervany [80] analyzed trust definitions used in 65 publications, and categorized them by conceptual type. They found that trust is often conceptualized as belief or intention in the IS literature, and that other conceptualizations are used less frequently. Moreover, the study lists corresponding operationalizations, thereby providing a valuable basis for the development of survey instruments. If trust is conceptualized as a belief, for example, it could be operationalized along characteristics of a trustee (e.g., benevolence, integrity, and competence), and possible measurement items could be, for example, "I believe that [the trustee] would act in my best interest," "If I required help, [the trustee] would do its best to help me," and "[the trustee] is interested in my well-being, not just its own" (e.g., [80]). Importantly, in the present paper we base trust measurements on observations of actual behavior (i.e., investment in the trust game). Hence, our measurement approach differs from that in many other IS trust studies, including those in the avatar domain, where trust measurement is either based on people's beliefs or intentions (survey and interview data; see [52, 86, 89, 93]) or on hypothetical scenarios (vignettes; see [103]). By measuring actual behavior, we are making a methodological contribution to the IS literature, thereby complementing the existing set of measurement approaches.

Future Work and Limitations

A possible avenue for future research would be to vary the degree of humanlikeness of avatars. For the present study, we used avatars that have a medium degree of humanlikeness (i.e., they are not completely simplistic cartoon-like characters, nor are they photorealistic portrayals; see Figure 2). It would be reasonable to assume that avatars with increasing humanlikeness should trigger brain activation patterns monotonically more similar to that of real humans. Research suggests, however, that such an assumption would be too simplistic, because as avatars approach photorealistic perfection but do not fully accomplish it, they cause humans to feel negative emotions (e.g., [24]) that neurologically resemble distrust reactions (e.g., [26]). This effect is referred to as the *uncanny valley effect* (e.g., [47, 77, 83]).

The participants in the present study had no information about whether or not the avatars represent actual humans, but were advised that the trustees would be unresponsive to their playing strategy, and that the character of each determines trustworthiness. Thus, another useful extension of our research could be to add an experimental condition in which participants are told that the avatars represent real humans (e.g., [5]). Such a manipulation could affect both trust behavior and activation in mentalizing brain areas, which would have far-reaching practical implications, for example, with respect to reputation mechanisms and the resulting collaboration effectiveness (for details, see [35, 42, 55, 88, 94]).

In our discussion of the theoretical rationale for H3, we indicated that learning is always accompanied by changes in the nervous system. Thus, learning can be studied at various analytical levels, ranging from observable behavioral changes to alterations in molecular structure. In the present study, our research focus was on the behavioral level. Consequently, future studies should adopt the neurobiological level to examine learning of another agent's trustworthiness (human or avatar). Voxel-based morphometry (VBM), for example, has recently been suggested in the IS literature as an appropriate tool for such studies [102], because this tool makes it possible to track gray matter density changes induced by learning processes within an individual (e.g., [30, 31]).

As is common in scientific research, the present fMRI study has limitations that should be taken into account. First, the interpretation of our empirical findings is based on a simple game-playing task in a controlled laboratory environment. Second, during the experiment, the participants were required to lie still and were restrained with pads to prevent motion during measurement sessions. Therefore, the experimental situation was relatively artificial, because in real-life conditions users typically sit in front of their computers. Accordingly, future IS avatar studies could measure users' psychophysiological responses (e.g., skin conductance, heart rate) to reduce experimental artificiality. Third, the present study investigates activation of mentalizing brain mechanisms by using static pictures of humans and avatars. Indeed, we found activation in a network spanning the MFC. However, inferring another actor's thoughts and intentions is not based solely on the act of processing static information. Motion (e.g., gestures) has also been demonstrated to enable inferences regarding another actor's mind (e.g., [42]). Future investigations could replicate our study using nonstatic stimulus material. Considering recent findings regarding the positive effects of both visual and behavioral realism of avatars on outcome variables such as affect-based trustworthiness (e.g., [12, 49, 92]), we hypothesize that, for an avatar, an increased level of humanlikeness that is induced by nonstatic information (e.g., gestures or animated facial expressions) could reduce the neurobiological differences between humans and avatars in mentalizing circuits [84]. In this context, one study [45] has already found that the mirror neuron system—which transforms observed actions into the neural representations of these actions—responds to both human and robotic actions, with no significant differences between these two agents.

Despite the limitations demonstrated by the range of important issues not addressed here, however, we believe that the present study contributes to a better understanding of trust and mentalizing in IS research, and holds significant value as a preliminary study for those that will build on the investigative line that we open here.

Concluding Comments

Trust and mentalizing are of essential value for social interaction [98]. From an evolutionary viewpoint, trust has been critical for survival over millions of years [21]. Most significantly, throughout history humans have faced risks related to trust, as trust in individuals who turned out not to be trustworthy has resulted in loss and even in death. Moreover, as a correlative condition for humans, being part of a social group has supported survival [76]. The division of basic human activities (e.g., one group member takes care of food acquisition, another provides protection from adversarial groups, and another cares for offspring) has resulted in a greater certainty of survival, while social groups without this structure have not achieved the same result. Most notably, the division of activities among group members implies cooperative behavior, which is strongly based on trust and mentalizing. Bearing these explanations in mind, it becomes clear that trust and mentalizing are intriguing—and essential—social phenomena that have evolved in order to secure the survival of humankind. The human face, in particular, has served as a reliable source of information for the prediction of another individual's trustworthiness, and for the consequent increased fitness for survival. Slowly, a trait for processing human facial information has been encoded in human genetic makeup. The recent emergence of computer-mediated forms of communication and cooperation, however, and most particularly the advent of avatars, promises to bring revolutionary changes in the organization and interaction of societies. Whatever new forms of social and economic interaction will confront society in the future, humans hold the ability to cope with the challenges, because evolution has endowed us with a remarkable, flexible brain.

Acknowledgments: The authors thank the guest editors, Ting-Peng Liang and Jan vom Brocke, as well as three anonymous reviewers for their excellent work in providing guidance on ways to improve the paper. Moreover, they are grateful to Cornelia Huber for her support in developing the avatars. They also appreciate the generous and visionary support of Novomind AG, Germany (www.novomind.com). They appreciate the generous support of Schindler Parent Meersburg who, through the Schindler Parent Distinguished Guest Lecturer for Marketing, supported René Riedl's work on NeuroIS and consumer neuroscience projects at Zeppelin University Friedrichshafen. Peter N.C. Mohr's work on this project was supported by the "Collaborative Research Center 649: Economic Risk" funded by the German Research Foundation (DFG). Fred D. Davis's work on this paper was partially supported by Sogang Business School's World Class University Project (R32-20002) funded by the Korean Research Foundation. The authors acknowledge the support of Hock Hai Teo and Hock Chuan Chan, who served as chairs of the track "Human–Computer Interaction" at the 2011 International Conference on Information Systems (ICIS, Shanghai, December 4–7), where an earlier version of this paper was presented. They also thank the conference participants who provided useful comments. Finally, they thank Deborah C. Nester for proofreading the manuscript.

References

1. Abraham, C.; Boudreau, M.-C.; Junglas, I.; and Watson, R. Enriching our theoretical repertoire: The role of evolutionary psychology in technology acceptance. *European Journal of Information Systems, 22,* 1 (2013), 56–75.

2. Abu-Akel, A. A neurobiological mapping of theory of mind. *Brain Research Reviews, 43,* 1 (2003), 29–40.

3. Adolphs, R. The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60* (2009), 693–716.

4. Amodio, D.M., and Frith, C.D. Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 4 (2006), 268–277.

5. Appel, J.; von der Pütten, A.M.; Krämer, N.C.; and Gratch, J. Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human–computer interaction. *Advances in Human–Computer Interaction,* (2012), article ID 324694, 1–10.

6. Bailenson, J.N., and Blascovich, J. Avatars. In W.S. Baindbridge (ed.), *Encyclopedia of Human–Computer Interaction.* Great Barrington, MA: Berkshire, 2004, pp. 64–68.

7. Bainbridge, W.S. The scientific research potential of virtual worlds. *Science, 317* (July 27, 2007), 472–476.

8. Bates, B., and Cleese, J. *The Human Face.* New York: DK, 2001.

9. Behrens, T.E.J.; Hunt, L.T.; and Rushworth, M.F.S. The computation of social behavior. *Science, 324* (May 29, 2009), 1160–1164.

10. Behrens, T.E.J.; Hunt, L.T.; Woolrich, M.W.; and Rushworth, M.F.S. Associative learning of social value. *Nature, 456* (November 13, 2008), 245–249.

11. Benbasat, I.; Dimoka, A.; Pavlou, P.A.; and Qiu, L. Incorporating social presence in the design of the anthropomorphic interface of recommendation agents: Insights from an fMRI study. In *Proceedings of the 31st International Conference on Information Systems.* Atlanta: Association for Information Systems, 2010, pp. 1–22.

12. Bente, G.; Rüggenberg, S.; Krämer, N.C.; and Eschenburg, F. Avatar-mediated networking: Increasing social presence and interpersonal trust in Net-based collaborations. *Human Communication Research, 34,* 2 (2008), 287–318.

13. Berg, J.; Dickhaut, J.; and McCabe, K. Trust, reciprocity, and social history. *Games and Economic Behavior, 10,* 1 (1995), 122–142.

14. Boaz, N.T., and Almquist, A.J. *Biological Anthropology: A Synthetic Approach to Human Evolution.* Upper Saddle River, NJ: Prentice Hall, 2001.

15. Bodamer, J. 1947. Die prosopagnosie. *Archiv für Psychiatrie und Nervenkrankheiten, 179,* 1–2 (1947), 6–53.

16. Buss, D.M. *Evolutionary Psychology: The New Science of the Mind.* Needham Heights, MA: Allyn & Bacon, 1999.

17. Cacioppo, J.T.; Berntson, G.G.; Sheridan, J.F.; and McClintock, M.K. Multilevel integrative analyses of human behavior: Social neuroscience and the complementing nature of social and biological approaches. *Psychological Bulletin, 126,* 6 (2000), 829–843.

18. Caldara, R.; Seghier, M.L.; Rossion, B.; Lazeyras, F.; Michel, C.; and Hauert, C.-A. The fusiform face area is tuned for curvilinear patterns with more high-contrasted elements in the upper part. *NeuroImage, 31,* 1 (2006), 313–319.

19. Carlson, J.R., and Zmud, R.W. Channel expansion theory and the experiential nature of media richness perceptions. *Academy of Management Journal, 42,* 2 (1999), 153–170.

20. Cartwright, J. *Evolution and Human Behavior: Darwinian Perspectives on Human Nature.* Cambridge: MIT Press, 2000.

21. Cosmides, L., and Tooby, J. Cognitive adaptations for social exchange. In J.H. Barkow, L. Cosmides, and J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* London: Oxford University Press, 1992, pp. 163–228.

22. Dahl, C.D.; Logothetis, N.K.; Bülthoff, H.H.; and Wallraven, C. The Thatcher illusion in humans and monkeys. *Proceedings of the Royal Society B: Biological Sciences, 277,* 1696 (2010), 2973–2981.

23. Darwin, C.R. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* London: John Murray, 1859.

24. Davidson, R.J. Affective neuroscience and psychophysiology: Toward a synthesis. *Psychophysiology, 40,* 5 (2003), 655–665.

25. Davis, A.; Murphy, J.; Owens, D.; Khazanchi, D.; and Zigurs, I. Avatars, people, and virtual worlds: Foundations for research in metaverses. *Journal of the Association for Information Systems, 10,* 2 (2009), 90–117.

26. Dimoka, A. What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly, 34,* 2 (2010), 373–396.

27. Dimoka, A.; Pavlou, P.A.; and Davis, F.D. NeuroIS: The potential of cognitive neuroscience for information systems research. *Information Systems Research, 22,* 4 (2011), 687–702.

28. Dimoka, A.; Banker, R.D.; Benbasat, I.; Davis, F.D.; Dennis, A.R.; Gefen, D.; Gupta, A.; Ischebeck, A.; Kenning, P.; Müller-Putz, G.; Pavlou, P.A.; Riedl, R.; vom Brocke, J.; and Weber, B. On the use of neurophysiological tools in IS research: Developing a research agenda for NeuroIS. *MIS Quarterly, 36,* 3 (2012), 679–702.

29. Donath, J. Virtually trustworthy. *Science, 317* (July 6, 2007), 53–54.

30. Draganski, B.; Gaser, C.; Busch, V.; Schuierer, G.; Bogdahn, U.; and May, A. Neuroplasticity: Changes in grey matter induced by training. *Nature, 427* (January 22, 2004), 311–312.

31. Driemeyer, J.; Boyke, J.; Gaser, C.; Büchel, C.; and May, A. Changes in gray matter induced by learning—Revisited. *PLoS ONE, 3,* 7 (2008), 1–5.

32. Dunbar, R.I.M. The social brain hypothesis. *Evolutionary Anthropology, 6,* 5 (1998), 178–190.

33. Ekman, P. *Emotion in the Human Face,* 2d ed. New York: Cambridge University Press, 1982.

34. Falloon, G. Using avatars and virtual environments in learning: What do they have to offer? *British Journal of Educational Technology, 41,* 1 (2010), 108–122.

35. Fehr, E. Human behaviour: Don't lose your reputation. *Nature, 432* (November 25, 2004), 449–450.

36. Fehr, E. On the economics and biology of trust. *Journal of the European Economic Association, 7,* 2–3 (2009), 235–266.

37. Fehr, E. Social preferences and the brain. In P.W. Glimcher, C.F. Camerer, E. Fehr, and R.A. Poldrack (eds.), *Neuroeconomics: Decision Making and the Brain.* Amsterdam: Academic Press, 2009, pp. 215–232.

38. Flinn, M.V. Evolution of stress response to social threat. In R.I.M. Dunbar and L. Barrett (eds.), *The Oxford Handbook of Evolutionary Psychology.* Oxford: Oxford University Press, 2007, pp. 273–295.

39. Franceschi, K.; Lee, R.M.; Zanakis, S.H.; and Hinds, D. Engaging group e-learning in virtual worlds. *Journal of Management Information Systems, 26,* 1 (Summer 2009), 73–100.

40. Frith, C.D., and Frith, U. The neural basis of mentalizing. *Neuron, 50,* 4 (2006), 531–534.

41. Frith, U., and Frith, C.D. Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358,* 1431 (2003), 459–473.

42. Frith, U., and Frith, C.D. The social brain: Allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365,* 1537 (2010), 165–176.

43. Galanxhi, H., and Nah, F.F.-H. Deception in cyberspace: A comparison of text-only vs. avatar-supported medium. *International Journal of Human–Computer Studies, 65,* 9 (2007), 770–783.

44. Gallagher, H.L., and Frith, C.D. Functional imaging of "theory of mind." *Trends in Cognitive Sciences, 7,* 2 (2003), 77–83.

45. Gazzola, V.; Rizzolatti, G.; Wicker, B.; and Keysers, C. The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage, 35,* 4 (2007), 1674–1684.

46. Gefen, D.; Benbasat, I.; and Pavlou, P.A. A research agenda for trust in online environments. *Journal of Management Information Systems, 24,* 4 (Spring 2008), 275–286.

47. Geller, T. Overcoming the uncanny valley. *IEEE Computer Graphics and Applications, 28,* 4 (2008), 11–17.

48. Goren, C.C.; Sarty, M.; and Wu, P.Y.K. Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics, 56,* 4 (1975), 544–549.

49. Groom, V.; Nass, C.; Chen, T.; Nielsen, A.; Scarborough, J.K.; and Robles, E. Evaluating the effects of behavioral realism in embodied agents. *International Journal of Human–Computer Studies, 67,* 10 (2009), 842–849.

50. Hamer, D. Rethinking behavior genetics. *Science, 298* (October 4, 2002), 71–72.

51. Haxby, J.V.; Gobbini, M.I.; Furey, M.L.; Ishai, A.; Schouten, J.L.; and Pietrini, P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293* (September 28, 2001), 2425–2430.

52. Hess, T.; Fuller, M.; and Campbell, D. Designing interfaces with social presence: Using vividness and extraversion to create social recommendation agents. *Journal of the Association for Information Systems, 10,* 12 (2009), 889–919.

53. Kanwisher, N.; McDermott, J.; and Chun, M.M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 11 (1997), 4302–4311.

54. Keeney, R.L. The value of Internet commerce to the customer. *Management Science, 45,* 4 (1999), 533–542.

55. King-Casas, B.; Tomlin, D.; Anen, C.; Camerer, C.F.; Quartz, S.R.; and Montague, P.R. Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308* (April 1, 2005), 78–83.

56. Kircher, T.; Blumel, I.; Marjoram, D.; Lataster, T.; Krabbendam, L.; Weber, J.; van Os, J.; and Krach, S. Online mentalising investigated with functional MRI. *Neuroscience Letters, 454,* 3 (2009), 176–181.

57. Knutson, B. Facial expression of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior, 20,* 3 (1996), 165–182.

58. Kock, N. The ape that used e-mail: Understanding e-communication behavior through evolution theory. *Communications of the Association for Information Systems, 5,* 3 (2001), 1–29.

59. Kock, N. The psychobiological model: Towards a new theory of computer-mediated communication based on Darwinian evolution. *Organization Science, 15,* 3 (2004), 327–348.

60. Kock, N. Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward e-communication tools. *IEEE Transactions on Professional Communication, 48,* 2 (2005), 117–130.

61. Kock, N. Information systems theorizing based on evolutionary psychology: An interdisciplinary review and theory integration framework. *MIS Quarterly, 33,* 2 (2009), 395–418.

62. Kock, N. *Evolutionary Psychology and Information Systems Research: A New Approach to Studying the Effects of Modern Technologies on Human Behavior.* New York: Springer, 2010.

63. Kock, N. Evolutionary psychology and information systems theorizing. In N. Kock (ed.), *Evolutionary Psychology and Information Systems Research: A New Approach to Studying the Effects of Modern Technologies on Human Behavior.* New York: Springer, 2010, pp. 3–37.

64. Kock, N., and Chatelain-Jardón, R. Four guiding principles for research on evolved information processing traits and technology-mediated task performance. *Journal of the Association for Information Systems, 12,* 10 (2011), 684–713.

65. Kock, N., and Hantula, D.A. Do we have e-collaboration genes? *International Journal of e-Collaboration, 1,* 1 (2005), i–ix.

66. Kohler, T.; Matzler, K.; and Fuller, J. Avatar-based innovation: Using virtual worlds for real-world innovation. *Technovation, 29,* 6–7 (2009), 395–407.

67. Kolb, B., and Whishaw, I.Q. Brain plasticity and behaviour. *Annual Review of Psychology, 49* (1998), 43–64.

68. Kolb, B., and Whishaw, I.Q. *Fundamentals of Human Neuropsychology,* 6th ed. New York: Worth, 2009.

69. Kolb, B.; Gibb, R.; and Robinson, T.E. Brain plasticity and behaviour. *Current Directions in Psychological Science, 12,* 1 (2003), 1–5.

70. Krach, S.; Hegel, F.; Wrede, B.; Sagerer, G.; Binkofski, F.; and Kircher, T. Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS ONE, 3,* 7 (2008), 1–11.

71. Krueger, F.; McCabe, K.; Moll, J.; Kriegeskorte, N.; Zahn, R.; Strenziok, M.; Heinecke, A.; and Grafman, J. Neural correlates of trust. *Proceedings of the National Academy of Sciences, 104,* 50 (2007), 20084–20089.

72. Lawrence, P., and Nohria, N. *Driven: How Human Nature Shapes Our Choices.* San Francisco: Jossey-Bass, 2002.

73. Lee, K.C., and Chung, N. A Web DSS approach to building an intelligent Internet shopping mall by integrating virtual reality and avatar. *Expert Systems with Applications, 28,* 2 (2005), 333–346.

74. LeGrand, R.; Mondloch, C.J.; Maurer, D.; and Brent, H.P. Neuroperception: Early visual experience and face processing. *Nature, 410* (April 19, 2001), 890.

75. Lieberman, M.D. Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58* (2007), 259–289.

76. Lieberman, M.D., and Eisenberger, N.I. Pains and pleasures of social life. *Science, 323* (February 13, 2009), 890–891.

77. MacDorman, K.F.; Green, R.D.; Ho, C.-C.; and Koch, C.T. Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior, 25,* 3 (2009), 695–710.

78. Mattay, V.S., and Goldberg, T.E. Imaging genetic influences in human brain function. *Current Opinion in Neurobiology, 14,* 2 (2004), 239–247.

79. McCabe, K.; Houser, D.; Ryan, L.; Smith, V.; and Trouard, T. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, 98,* 20 (2001), 11832–11835.

80. McKnight, D.H., and Chervany, N.L. Trust and distrust definitions: One bite at a time. In R. Falcone, M. Singh, and Y.-H. Tan (eds.), *Trust in Cyper-Societies.* Berlin: Springer, 2001, pp. 27–54.

81. McMeill, D. *The Face: A Natural History.* Boston: Little, Brown, 1998.

82. Miller, G. The promise of parallel universes. *Science, 317* (September 7, 2007), 1341–1343.

83. Mori, M. Bukimi no tani [The uncanny valley]. *Energy, 7,* 4 (1970), 33–35.

84. Morris, J.P.; Pelphrey, K.A.; and McCarthy, G. Regional brain activation evoked when approaching a virtual human on a virtual walk. *Journal of Cognitive Neuroscience, 17,* 11 (2005), 1744–1752.

85. Moser, E.; Derntl, B.; Robinson, S.; Fink, B.; Gur, R.C.; and Grammer, K. Amygdala activation at 3T in response to human and avatar facial expressions of emotions. *Journal of Neuroscience Methods, 161,* 1 (2007), 126–133.

86. Mueller, J.; Hutter, K.; Fueller, J.; and Matzler, K. Virtual worlds as knowledge management platform: A practice-perspective. *Information Systems Journal, 21,* 6 (2011), 470–501.

87. Nowak, K.L., and Rauh, C. The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *Journal of Computer-Mediated Communication, 11,* 1 (2005), 153–178.

88. Nowak, M.A., and Sigmund, K. Evolution of indirect reciprocity. *Nature, 437* (October 27, 2005), 1291–1298.

89. Nunamaker, J.F., Jr.; Derrick, D.C.; Elkins, A.C.; Burgoon, J.K.; and Patton, M.W. Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems, 28,* 1 (Summer 2011), 17–48.

90. Oosterhof, N.N., and Todorov, A. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105,* 32 (2008), 11087–11092.

91. Premack, D., and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1,* 4 (1978), 515–526.

92. Qiu, L., and Benbasat, I. Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International Journal of Human–Computer Interaction, 19,* 1 (2005), 75–94.

93. Qiu, L., and Benbasat, I. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems, 25,* 4 (Spring 2009), 145–181.

94. Raub, W., and Weesie, J. Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology, 96,* 3 (1990), 626–654.

95. Rendall, D.; Notman, H.; and Vokey, J.R. Homologizing the mind: Frodoian or fodorian psychology? In R.I.M. Dunbar and L. Barrett (eds.), *The Oxford Handbook of Evolutionary Psychology.* Oxford: Oxford University Press, 2007, pp. 59–70.

96. Rescorla, R.A., and Wagner, A.R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy (eds.), *Classical Conditioning II: Current Research and Theory.* New York: Appleton-Century-Crofts, 1972, pp. 64–99.

97. Riedl, R. Zum Erkenntnispotenzial der kognitiven Neurowissenschaften für die Wirtschaftsinformatik: Überlegungen anhand exemplarischer Anwendungen [On the potential of cognitive neuroscience for business informatics: Thoughts based on exemplary applications]. *NeuroPsychoEconomics, 4,* 1 (2009), 32–44.

98. Riedl, R., and Javor, A. The biology of trust: Integrating evidence from genetics, endocrinology and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics, 5,* 2 (2012), 63–91.

99. Riedl, R.; Hubert, M.; and Kenning, P. Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *MIS Quarterly, 34,* 2 (2010), 397–428.

100. Riedl, R.; Kindermann, H.; Auinger, A.; and Javor, A. Technostress from a neurobiological perspective: System breakdown increases the stress hormone cortisol in computer users. *Business & Information Systems Engineering, 4,* 2 (2012), 61–69.

101. Riedl, R.; Mohr, P.N.C.; Kenning, P.H.; Davis, F.D.; and Heekeren, H.R. Trusting humans and avatars: Behavioral and neural evidence. In *Proceedings of the 32nd International Conference on Information Systems.* Atlanta: Association for Information Systems, 2011, pp. 1–23.

102. Riedl, R.; Banker, R.D.; Benbasat, I.; Davis, F.D.; Dennis, A.R.; Dimoka, A.; Gefen, D.; Gupta, A.; Ischebeck, A.; Kenning, P.; Müller-Putz, G.; Pavlou, P.A.; Straub, D.W.;

vom Brocke, J.; and Weber, B. On the Foundations of NeuroIS: Reflections on the Gmunden Retreat 2009. *Communications of the Association for Information Systems, 27* (2010), 243–264.

103. Robert, L.P., Jr.; Dennis, A.R.; and Hung, Y.C. Individual swift trust and knowledge trust in face-to-face and virtual team members. *Journal of Management Information Systems, 26,* 2 (Fall 2009), 241–279.

104. Rossion, B.; Hanseeuw, B.; and Dricot, L. Defining face perception areas in the human brain: A large-scale factorial fMRI face localizer analysis. *Brain and Cognition, 79,* 2 (2012), 138–157.

105. Rudoy, J.D., and Paller, K.A. Who can you trust? Behavioral and neural differences between perceptual and memory-based influences. *Frontiers in Human Neuroscience, 3* (2009), 1–6.

106. Schlicht, E.J.; Shimojo, S.; Camerer, C.F.; Battaglia, P.; and Nakayama, K. Human wagering behavior depends on opponents' faces. *PLoS ONE, 5,* 7 (2010), 1–10.

107. Schultze, U. Embodiment and presence in virtual worlds: A review. *Journal of Information Technology, 25,* 4 (2010), 434–449.

108. Schultze, U., and Orlikowski, W.J. Virtual worlds: A performative perspective on globally distributed, immersive work. *Information Systems Research, 21,* 4 (2010), 810–821.

109. Singer, T. Understanding others: Brain mechanisms of theory of mind and empathy. In P.W. Glimcher, C.F. Camerer, E. Fehr, and R.A. Poldrack (eds.), *Neuroeconomics: Decision Making and the Brain.* Amsterdam: Academic Press, 2009, pp. 251–268.

110. Sparrow, B.; Liu, J.; and Wegner, D.M. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333* (August 5, 2011), 776–778.

111. Sripada, C.S.; Angstadt, M.; Banks, S.; Nathan, P.J.; Liberzon, I.; and Phan, K.L. Functional neuroimaging of mentalizing during the trust game in social anxiety disorder. *NeuroReport, 20,* 11 (2009), 984–989.

112. Suh, K.-S.; Kim, H.; and Suh, E.K. What if your avatar looks like you? Dual-congruity perspectives for avatar use. *MIS Quarterly, 35,* 3 (2011), 711–729.

113. Sutton, R.S., and Barto, A.G. *Reinforcement Learning: An Introduction.* Cambridge: MIT Press, 1998.

114. Thompson, P.; Cannon, T.D.; Narr, K.L.; van Erp, T.; Poutanen, V.-P.; Huttunen, M.; Lönnqvist, J.; Standertskjöld-Nordenstam, C.-G.; Kaprio, J.; Khaledy, M.; Dail, R.; Zoumalan, C.I.; and Toga, A.W. Genetic influences on brain structure. *Nature Neuroscience, 4,* 12 (2001), 1253–1258.

115. Todd, P.M., and Gigerenzer, G. Mechanisms of ecological rationality: Heuristics and environments that make us smart. In R.I.M. Dunbar and L. Barrett (eds.), *The Oxford Handbook of Evolutionary Psychology.* Oxford: Oxford University Press, 2007, pp. 197–210.

116. Todorov, A. Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences, 1124* (2008), 208–224.

117. Todorov, A.; Said, C.P.; Engell, A.D.; and Oosterhof, N.N. Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12,* 12 (2008), 455–460.

118. Vom Brocke, J.; Riedl, R.; and Léger, P.-M. Application strategies for neuroscience in information systems design science research. *Journal of Computer Information Systems, 53,* 3 (2013), 1–13.

119. Vuilleumier, P.; Armony, J.L.; Driver, J.; and Dolan, R.J. Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience, 6,* 6 (2003), 624–631.

120. Wei, X.; Zhu, Z.; Yin, L.; and Ji, Q. A real time face tracking and animation system. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* Los Alamitos, CA: IEEE Compuater Society, 2004, pp. 1–8.

121. Willis, J., and Todorov, A. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17,* 7 (2006), 592–598.

122. Winston, J.S.; Strange, B.A.; O'Doherty, J.; and Dolan, R.J. Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5,* 3 (2002), 277–283.

Appendix

We acquired 2 runs of 690 functional T2*-weighted echoplanar images (EPI) [TR [repetition time], 2 s; echo time (TE), 40 ms; flip angle, 90°; field of view, 256 mm; matrix, 64 × 64 mm; 26 axial slices approximately parallel to the bicommissural plane; slice thickness, 4 mm]. In addition, for registration purposes, a high-resolution T1-weighted structural image (MPRAGE) was acquired from each participant [TR, 20 ms; TE, 5 ms; flip angle, 30°; 179 sagittal slices; voxel size, 1 × 1 × 1 mm]. Initial analysis was performed using the FSL toolbox from the Oxford Centre for fMRI of the Brain (www.fmrib.ox.ac.uk/fsl/). The image time-course was first realigned to compensate for small head movements. Data were spatially smoothed using an 8 mm full-width-half-maximum Gaussian kernel and were temporally filtered using a high-pass temporal filter (with sigma = 100 s). Registration was conducted through a two-step procedure, whereby EPI images were first registered to the MPRAGE structural image and then to standard MNI (Montreal Neurological Institute) space (MNI152_T1_2mm_brain), using 7 parameters for the first registration step and 12 parameters for the second. Statistical analyses were performed in native space, with the statistical maps normalized to standard space prior to higher-level analysis. Statistical analysis of functional data was performed using a multilevel approach implementing a mixed-effects model treating participants as a random effect. This was initially performed separately for each participant's concatenated runs. Regressors-of-interest were created by convolving a rectangular function representing stimulus duration times with a canonical (double-gamma) hemodynamic response function. Time-series statistical analysis was carried out using FILM (FMRIB's Improved Linear Model) with local autocorrelation correction. The functional analysis was based on two regressors-of-interest and four regressors-of-no-interest. Two binary regressors modeled the decision phase. The first regressor represents the decision in the human condition, whereas the second regressor models the decision in the avatar condition. The durations of the two regressors correspond to the decision time, which was three seconds in all cases. The final four regressors-of-no-interest modeled the trustworthiness ratings for humans and avatars as well as the outcome presentations for both conditions. On the group level (second level of analysis), we integrated the results from the single-subject level (first level), again applying a general linear model. One binary regressor modeled a constant effect of all the first-level parameter estimates on the group level.

PLOS | ONE

RESEARCH ARTICLE

Parkinson Patients' Initial Trust in Avatars: Theory and Evidence

Andrija Javor[1]*, Gerhard Ransmayr[1], Walter Struhal[1], René Riedl[2,3]

1 Department of Neurology II, Kepler University Clinic, Johannes Kepler University, Linz, Austria, **2** School of Management, University of Applied Sciences Upper Austria, Steyr, Austria, **3** Department of Business Informatics–Information Engineering, Johannes Kepler University, Linz, Austria

* javor@gmx.at

Citation: Javor A, Ransmayr G, Struhal W, Riedl R (2016) Parkinson Patients' Initial Trust in Avatars: Theory and Evidence. PLoS ONE 11(11): e0165998. doi:10.1371/journal.pone.0165998

Editor: Stefano Federici, Universita degli Studi di Perugia, ITALY

Received: December 20, 2015

Accepted: October 23, 2016

Published: November 7, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by a research grant from the Internet Foundation Austria (Call 8 – No. 688), https://www.netidee.at. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Parkinson's disease (PD) is a neurodegenerative disease that affects the motor system and cognitive and behavioral functions. Due to these impairments, PD patients also have problems in using the computer. However, using computers and the Internet could help these patients to overcome social isolation and enhance information search. Specifically, avatars (defined as virtual representations of humans) are increasingly used in online environments to enhance human-computer interaction by simulating face-to-face interaction. Our laboratory experiment investigated how PD patients behave in a trust game played with human and avatar counterparts, and we compared this behavior to the behavior of age, income, education and gender matched healthy controls. The results of our study show that PD patients trust avatar faces significantly more than human faces. Moreover, there was no significant difference between initial trust of PD patients and healthy controls in avatar faces, while PD patients trusted human faces significantly less than healthy controls. Our data suggests that PD patients' interaction with avatars may constitute an effective way of communication in situations in which trust is required (e.g., a physician recommends intake of medication). We discuss the implications of these results for several areas of human-computer interaction and neurological research.

Introduction

Parkinson's disease (PD) is a common, chronic progressive disease with a median annual incidence of 14 per 100.000 increasing to 160 per 100.000 in the age group over 65 [1]. Typical symptoms include bradykinesia, resting tremor, rigidity, and impaired postural reflexes [2]. In addition to these 'cardinal motor symptoms', non-motor symptoms occur, even in the early stages of the disease [3]. Cognitive and behavioral disturbances include deficits in the following domains: attention, memory, visuospatial functions and decision-making [4], as well as impairment of face recognition (e.g. [5]), risk-taking (e.g. [6]), trust [7], and ability to explain and predict other people's behavior (referred to as Theory-of-Mind or mentalizing, e.g., [8, 9]). The underlying pathology is a dopaminergic cell loss in the substantia nigra pars compacta. In addition, neuronal loss is reported in the following brain areas: the dorsal motor nucleus

(DMN) of the vagus, the olfactory bulb, the caudal raphe nuclei and the locus ceruleus [10]. Also, PD leads to the occurrence of Lewy bodies and neurochemical alterations in the brainstem, limbic system, diencephalic and basal brain structures, and in advanced stages, the neocortex [10]. Importantly, all mentioned deficits and impairments may negatively affect the social life of PD patients. In fact, cognitive impairment has been shown to predict quality of life in PD patients and often troubles patients more than their motor symptoms [11].

Patients suffering from all kinds of illnesses are frequently using information technology (IT) for private and disease specific information search and communication [12]. However, interaction with computers and the Internet may provoke a perception of complexity and uncertainty [13]. Evidence shows that nearly 80% of computer users with PD, report significant, severe, or highly severe difficulties using a computer [14]. Underlying factors identified in this study are inertia, muscle stiffness, tremor, and issues related to the use of input devices, among other aspects related to ergonomics. To overcome impairments, the design of human-computer interfaces for disease populations has become a major topic in human-computer interaction (HCI) research [15]. While the effects of motor symptoms on PD patients' interaction with computers has received considerable attention in the HCI domain (e.g. [16]), behavioral symptoms and their effects have largely been neglected. However, there is growing evidence that computer and Internet use may help to provide essential information, reduce loneliness, and increase social contact among older adults in assisted and independent living communities [17]. This finding is of paramount importance for PD patients, because they frequently suffer from social isolation [18].

Avatars and social behavior in HCI

In the HCI domain, avatars are defined as user-created digital representations symbolizing the user's presence in a virtual environment [19]. In addition, artificial faces and characters might be part of a computer interface. In this case, characters are created by systems engineers and controlled by computer algorithms (and not by a human user). Characters controlled by algorithms are referred to as agents [19]. Examples for agents are virtual salespersons in an online shop or digital tutors in e-learning environments. To avoid complex sentence structures, we only use the term avatar in the present paper. However, we note that in this paper we refer to both avatars and agents.

Social information processing theory [20] postulates that users engage in interpersonal processes online in order to attain interpersonal goals. A major goal of communicators is the reduction of uncertainty about others [21]. Visual appearance plays a significant role in the formation of first impressions of strangers, and this visual and nonverbal information is processed to form a first impression within seconds [22]. Thus, provision of visual information may significantly reduce uncertainty in social interaction. Importantly, first impressions are frequently very resistant to change [23]. It follows that first impressions significantly affect future behavior of communicators in the social interaction process. In online environments, visual information is often provided based on static information (e.g., by showing a communicator's face or body [24]). Importantly, avatars are able to induce trust and reduce uncertainty despite being presented statically, or as an animation [25, 26].

There is vast evidence on how healthy humans interact with avatars. Several studies based on healthy subjects have demonstrated that avatars are often perceived as social agents, and thus avatars may socially influence humans. Hence, humans interacting with avatars have an experience of being with another person [27] and are willing to disclose information that is considered highly personal, even when avatars are low in visual realism [28]. Moreover, recent evidence indicates that avatars are trusted to a similar extent as humans [29]. Thus, avatars help to overcome perceptions of uncertainty in computerized environments by imitating face-to-face interaction [30].

In addition to research on healthy subjects, there have been first studies to investigate how patients suffering from psychiatric diseases, such as schizophrenia (e.g. [31]) and autism (e.g. [32]), interact with avatars. However, to the best of our knowledge, neither does a peer-reviewed scientific journal article focus on neurological patients' interaction with avatars, nor does peer-reviewed research at the nexus of avatars and PD exist. This finding is surprising, because avatars are used in (i) computer systems that are specifically designed to evaluate PD patients [33], (ii) assistive technologies [34], and (iii) neurorehabilitation systems for PD patients [35].

To close this significant research gap, we investigate the trust behavior of PD patients during their interaction with avatar and human faces in a trust game context. In this game, originally developed in behavioral economics [36], an investor (also referred to as trustor) receives an initial monetary endowment (e.g. €10). The investor decides how much of this initial endowment he or she would like to send to another player in the game, the trustee. This amount is then multiplied by the experimenter (e.g. €10 x 6 = €60). In the next move, the trustee decides whether to send money back to the investor (and, if so, how much). The amount of the investor's transfer is a behavioral measure of initial trust [36–38].

In the next section, we give an overview of (1) the pathoanatomical basis of major behavioral symptoms in PD and (2) the vast literature on the neurobiology of trust in both interpersonal and human-avatar interaction.

Behavioral symptoms in PD

Behavioral and cognitive neurology, a field that studies behavioral symptoms and cognitive functions in patients suffering from neurological disorders, has identified several impairments in PD. Cognitive symptoms of PD primarily affect executive functions [39]. It is still a matter of research why some PD patients develop cognitive impairments in addition to the cardinal motor symptoms in early stages of the disease, while others do not, or only at a very late stage of the disease. The reason for this observation might be that motor symptoms and cognitive deficits have distinct underlying anatomical bases [40]. In the following, we review the PD literature on pathological changes in brain regions relevant to trust, and we relate these regions to cognitive and behavioral impairments in PD.

First, PD pathologically affects the limbic system, particularly the amygdala and the thalamus (e.g. [41, 42]). A network of brain structures including the amygdala underlies the ability of face recognition [43]. This explains why PD patients have been shown to have problems in decoding facial expressions and hence perform significantly worse in recognizing sadness, anger, and disgust in human faces when compared to healthy controls (e.g. [44]).

Second, in PD, the loss of dopaminergic neurons in the substantia nigra leads to pathological functioning of striatal structures [45] and to impairments in the meso-cortico-limbic dopaminergic system, including the ventral tegmental area, ventral striatum, and medial orbitofrontal cortex [46]. Moreover, a network of brain structures that involves multiple cortical and subcortical regions (prefrontal, parietal, limbic) constitutes the neural basis of decision-making under risk or uncertainty [47]. This network overlaps in critical parts with the reward system (striatum and frontal cortical regions, e.g. [48]), and therefore explains altered risk processing in PD patients. Drug naïve patients are typically risk averse, while dopaminergic treatment makes these patients more prone to risk-seeking behavior (for a review, see [6]).

Imaging data also suggests that frontal cortical regions are affected by PD (e.g. [49]), and frontal lobe functions are disturbed in PD in about 30% of the patients [3]. Impairments in mentalizing (Theory-of-Mind) have also been reported for PD patients (for a review, see [50]), and these disturbances are mainly caused by frontal cortex dysfunction, because the frontal cortex is considered a key region for the neural implementation of mentalizing (e.g. [8]). A recent review of 13 studies focusing on mentalizing in PD concludes that (i) deficits are common and (ii) those deficits probably play an important role in daily living and affects quality of life in these patients [9].

To sum up, we have briefly reviewed evidence on the pathological changes of the limbic system, the basal ganglia, and the frontal cortex in PD. Importantly, human trust relies on all of these brain regions (for a review, see [51]), and PD patients have recently been shown to have lower trust in other humans than healthy controls [7].

Next, we provide a short overview of the literature on the neurobiology of trust.

The neurobiology of trust

Trust is a fundamental prerequisite for human relationships, both in private and public life, and thus is essential for the functioning of society in general [52]. It is important to differentiate trust behavior from its antecedents (beliefs about trust, behavioral intentions, and trust disposition) as these concepts are frequently mixed up in the literature [53]. Following the behavioral intentions concept, Rousseau et al. [54] defined trust as a psychological state comprising the intention to accept vulnerability based on positive expectations about the actions of another party (the trustee). Formally, trust is therefore the subjective assessment of the probability that someone is trustworthy [55]. Trust *behavior* has been defined as trusting acts influenced by beliefs about another person's trustworthiness and risk preference [56]. Initial trust is trust a person has in a stranger before having any interaction experience with that person. This kind of trust plays an important role in online environments, as users often have little common history or may not share the same cultural background [57]. Knowledge-based trust, in contrast, is trust that emerges through engagement with the other person. While initial trust strongly depends on the characteristics of another person (e.g. face), knowledge-based trust primarily develops from another person's reciprocal behavior. One person, the trustor, is acting in a way that makes him vulnerable to the actions of another person, the trustee, in the sense that trust might not be reciprocated. Due to a social preference referred to as"betrayal aversion", humans generally take risks less willingly when the cause of uncertainty is another person, and this fact differentiates trust from non-social risk-taking (e.g. [58, 59]).

Trust has become a major research topic in various scientific disciplines, including economics, neuroscience, medicine and information systems research (e.g. [58, 60, 61]). Neuroscience has used several methodological approaches (e.g. functional brain imaging) to detect the neural correlates of trust. In the following, we summarize key findings from an anatomical point of view and, in order to enhance understandability from a non-neuroscience perspective, group several smaller brain areas into three main regions that were found to be significantly activated in trust situations, namely the limbic system, basal ganglia, and the frontal cortex (for more comprehensive reviews, see [51, 58, 62]).

The limbic system is a term used for several brain regions, including the amygdala and the hippocampus that are strongly related to emotions and memory [63]. The amygdala is active during the assessment of the trustworthiness of human faces (e.g. [64–66]), a fact that holds particularly true for the assessment of untrustworthy faces [43, 66]. A study based on the trust game found that patients with unilateral amygdala damage showed more trust than healthy controls [67]. This finding suggests that the amygdala plays a crucial role in trust behavior between humans.

The basal ganglia are a subset of subcortical nuclei that are interconnected with several cortical areas (especially the frontal cortex), as well as the thalamus and the brainstem. The basal ganglia have important functions for the motor system, but also for learning, cognition, and behavior (e.g. [68]). One part of the basal ganglia is the striatum, which is active in trust situations because the striatum is associated with reward processing and reward anticipation (e.g. [64–69, 70]). The caudate nucleus, which is also part of the basal ganglia, has also been shown to be active in trust game studies [65].

Theory-of-mind brain regions are needed to predict another person's trustworthiness based on anticipation of intentions and future behavior. The outcome of this process is known as 'calculative trust' [71]. Theory-of-mind regions that have been shown to be active in trust behavior are the paracingulate cortex [72] and the medial prefrontal cortex [73]; we group both regions together and refer to them as frontal cortical regions in our paper. These regions seem to have distinct functions in trust behavior, as the paracingulate cortex has a role in building a trust relationship by inferring another person's intentions to predict subsequent behavior, and the medial prefrontal cortex has implications for decisions and choices based on calculative expectations of what others will do [74].

Hypotheses

We provided an overview of the most important aspects of the biological foundations of social behavior deficits in PD and the neurobiology of trust. It becomes apparent that there is a significant overlap between pathoanatomical changes in PD and the anatomical basis of trust. We visualize this overlap in Fig 1.

In order to develop a theoretical basis for our hypotheses that describe how PD patients are expected to behave in a trust situation with avatars, we summarize evidence on PD patients' trust in human faces and healthy controls' trust in avatars and human faces in Fig 2. Based on

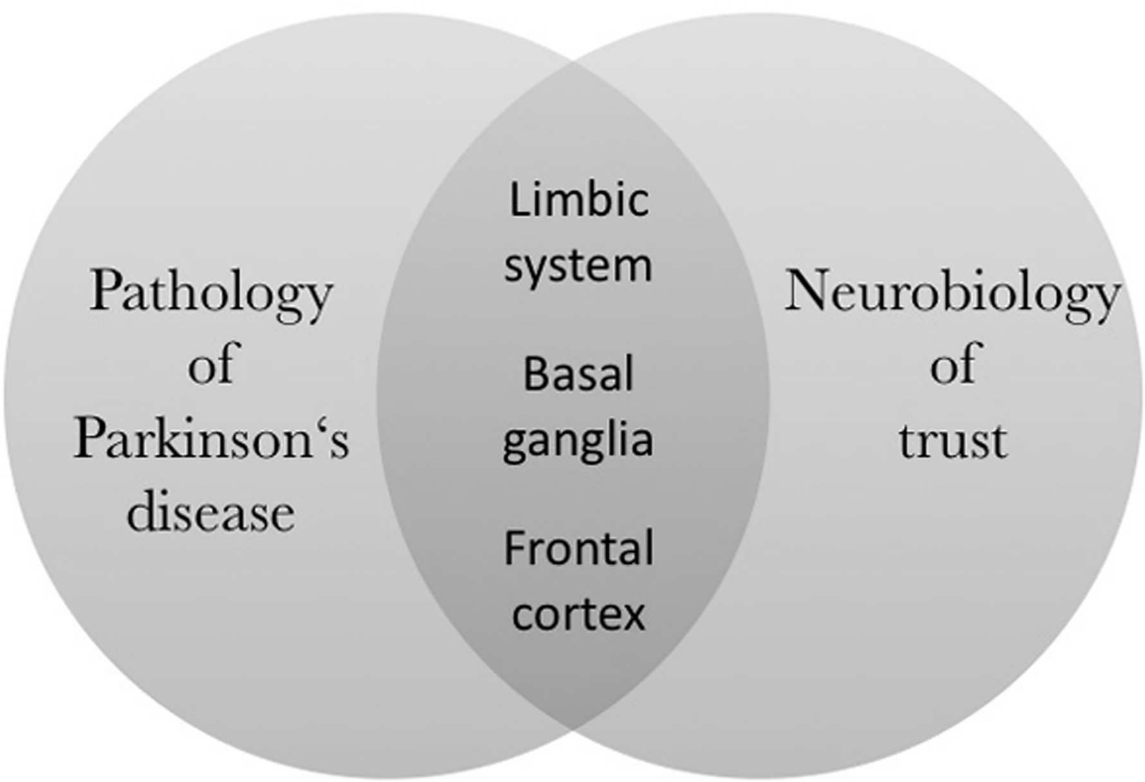


Fig 1. Brain regions related to trust and affected by PD.

doi:10.1371/journal.pone.0165998.g001

		Investor	
		PD patients	*Healthy subjects*
Trustee	*Avatar*	**This study**	Riedl et al., 2011 and 2014
	Human	Javor et al., 2015	Johnson & Mislin, 2011 (meta-analysis)

Fig 2. Visualization of current knowledge of trust research and the contribution of the present study.

doi:10.1371/journal.pone.0165998.g002

this conceptualization, we identify two significant gaps in literature. In the following, we develop hypotheses for these two gaps. Based on our preceding discussion of neuroscience research, we argue as follows:

1. Trust is mainly represented in three brain regions, namely the limbic system, the basal ganglia, and frontal cortical regions; PD affects all these regions. It has been shown that PD patients have lower initial trust in simulated face-to-face interactions compared to healthy controls [7].
2. There is evidence that mentalizing brain regions are more active during the evaluation of the trustworthiness of human faces if compared to avatar faces [29]. Furthermore, avatar faces elicit less activation in the limbic system if compared to human faces [26].
3. One could argue that even though PD patients show lower initial trust in simulated face-to-face interactions than healthy controls [7], these patients have less or even no trust deficit when interacting with avatars, because brain regions that are regulating trust behavior (e.g. medial frontal cortex and the amygdala) are less active in the interaction with avatars compared to interaction with humans [26, 29]. Therefore, impairments in these regions caused by PD should have less impact on trust behavior towards avatars.

Following this line of argumentation, we formulate the following hypotheses that we tested in a laboratory experiment:

H1: PD patients show higher initial trust in avatar faces than in human faces.

H2: PD patients show similar initial trust in avatar faces if compared to healthy controls.

Methods

This study was approved by the ethics committee of Upper Austria. Informed written consent was obtained from all participants prior to the experiment.

Participants

Forty right-handed, Caucasian participants were recruited for the study. Twenty of the subjects were patients diagnosed with PD according to UK PDS Brain Bank criteria [75], an instrument with a diagnostic accuracy of 90% [76]. The patients were recruited at an outpatient clinic specialized in movement disorders (mean age 72.35 years, SD 9.16, equally gendered). The same patient population was studied before and more clinically relevant variables are available elsewhere (see [7]). This study is part of an extended project to explore trust in PD. All patients were clinically stable under treatment prior to the experiment. Healthy controls had a mean age of 68.4 years (SD 10, also equally gendered), no history of neurological disease, and a neurological status with no pathological findings. There was no significant difference between PD patients and healthy controls with respect to: age (two-sided unpaired t-test; $t = 1.3$, $p = 0.20$), income (classified in 6 categories; median patients: 4; median controls: 3; Mann-Whitney-U-test; $z = -0.86$; $p = 0.39$), education (two-sided unpaired t-test; mean education patients: 10.9 years, SD 4.02; mean education controls: 11.3 years, SD 5.11; $t = -0.28$; $p = 0.79$), and religion (all participants were Roman-Catholic); note that these factors might affect trust behavior [77].

We screened for psychiatric co-morbidities like dementia in all patients and controls, as well as for impulsive-compulsive disorders, and/or apathy in all patients and excluded those meeting pathological criteria. As screening tools we used the complete version of the Patient Health Questionnaire (PHQ-D, [78]), the Questionnaire for Impulsive-Compulsive Disorders (QUIP, [79]), the Mini Mental State Examination (MMSE, [80]), and the Lille Apathy Rating Scale (LARS, [81]). To quantify the severity of symptoms in PD we also used the Unified Parkinson's Disease Rating Scale (UPDRS, [82, 83]), a scale commonly used in clinical practice. The mean UPDRS-III score of patients under treatment was 19.85 (SD 10.59). For a summary of the characteristics of our study population see Table 1.

Experimental Procedure

We designed a behavioral experiment to test our hypotheses. When subjects entered the laboratory for the experiment, they received written instructions explaining the rules and payoff structure for the experiment (the instructions are available from the first author on request). The experimenter subsequently checked whether the subjects understood the instructions by going through several hypothetical examples. All subjects correctly answered the control questions.

Table 1. Characteristics of the patient and control groups.

Characteristics	*PD patients*	*Healthy controls*
Number of subjects (n)	20	20
Female, n (%)	10 (50)	10 (50)
Age (years)	72.35±9.16	68.4±10
Education (years)	10.9±4.03	11.3±5.11
MMSE	28.4±1.6	28.95±0.99

Abbreviations: MMSE: Mini Mental State Examination. Except where indicated, values are means plus/minus standard deviation.

doi:10.1371/journal.pone.0165998.t001

We measured trust using a version of the original trust game [36]. The trust game was developed to measure trust as actual behavior of players in an economic exchange game. In this game, one player (the trustor), has an initial endowment of x monetary units. First, the trustor decides whether to keep his or her endowment, which ends the game, or to send (a part of) it to a second player (the trustee). The trustee observes the trustor's action and, if money was sent, decides whether to keep the amount or share (some of) it with the trustor. The experimenter multiplies the trustor's transfer, so that both players are better off collectively if the trustor transfers money and the trustee sends back a sufficient amount. In the trust game, the amount sent by the trustor is used as a behavioral measure for trust [36–38].

Our version of the game focused on the trustor's behavior, whose role was played by the participants on a computer. The participants were told that they would be playing against human beings presented in the form of face photographs or avatar faces, but actually played against an automatic and randomized computer-generated strategy that simulated a trustee's behavior (see Fig 3 for examples of human faces and avatar faces which we used as stimulus material).

The initial endowment of the participants was €10, and they were told that they would be playing with real money and that the recompense for participating in the experiment would reflect their gain in the game. The participants could decide to send any amount to the trustee but had to use whole numbers between €0 and €10. This amount sent by the participants was used to operationalize initial trust. The trustee was illustrated with 16 human face images (taken from an established face database [84, 85]) and 16 avatar faces (taken from the stimulus material from published research [29, 86]). The selection of the human and avatar stimuli ensured the same face trustworthiness and gender in both groups. None of the human faces or avatar faces were known or previously seen by the participants. It follows that participants had no interaction experience with the stimulus material. In total, each subject made 32 decisions in the experiment, and the order of face presentation was completely random. It follows that no participant interacted with the same trustee twice. We used the monetary amount sent by each participant as a measure of trust behavior. Fig 4 illustrates our version of the trust game. We included a gambling task (Game-of-Dice task, for a detailed description of the game see [4]) in our experimental design, because non-social risky decision-making is considered a standard control condition in trust research (e.g. [87]).

Statistical Analysis

We used SPSS® (Version 20) to perform statistical analyses. Descriptive statistics were run to illustrate the demographic characteristics of the sample and to describe disease specific information about the patient group (e.g. UPDRS-II scores). A Kolmogorov-Smirnov test revealed that game behavior of both experimental groups conformed to normality and a Levene's test of equality assured that there was homogeneity of variances between groups. Unpaired sample two-sided t-tests were performed to compare game behavior between groups and paired sample two-sided t-tests to compare within-group behavior between the human face and avatar face conditions.

Results

H1 states that PD patients show higher initial trust in avatar faces than in human faces. H1 is supported by our data. PD patients invested significantly more in trust games played against avatar faces versus games played against human faces (mean avatar faces: €4.91, SD 1.55; mean human faces: €3.43, SD 2.00; two-sided paired sample t-test; $t = 2.62$; $p = 0.017$, see Fig 5). In contrast, healthy controls' mean investment amounts did not differ significantly in trust games

Fig 3. Examples of stimulus material used in this study.

doi:10.1371/journal.pone.0165998.g003

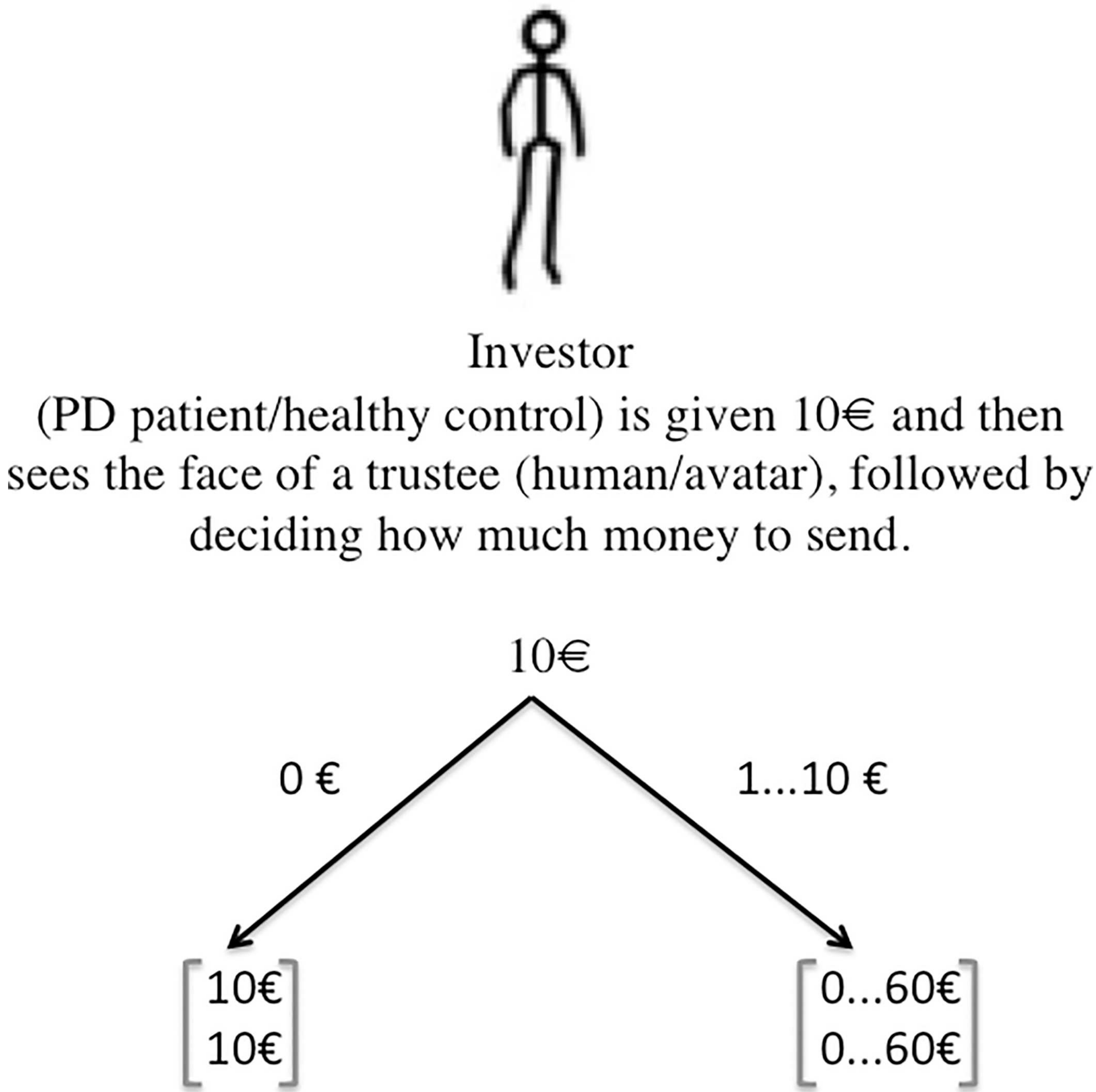


Fig 4. Illustration of our version of the trust game. Notes: The upper value in the square brackets indicates the investor's payoff (any amount between €0 and 60), the lower value the trustee's payoff depending on the investor's first move (investment of €0 or any amount between €1 and 10). When the investor is not sending any money to the trustee (€0) the payoff is €10 for both players. In any other case, the investor's payoff is dependent on the trustee's willingness to send some money back. Whereas the investor's payoff determined the overall gain of the participants in our study, the trustee's payoff was not paid out, as this role was not played by participants, but was part of the computerized experiment (see Experimental procedure).

doi:10.1371/journal.pone.0165998.g004

against avatar faces versus human faces (mean avatars: €4.94, SD 1.55; mean human faces: €5.53, SD 1.56; two-sided paired sample t-test; $t = -1.53$; $p = 0.141$).

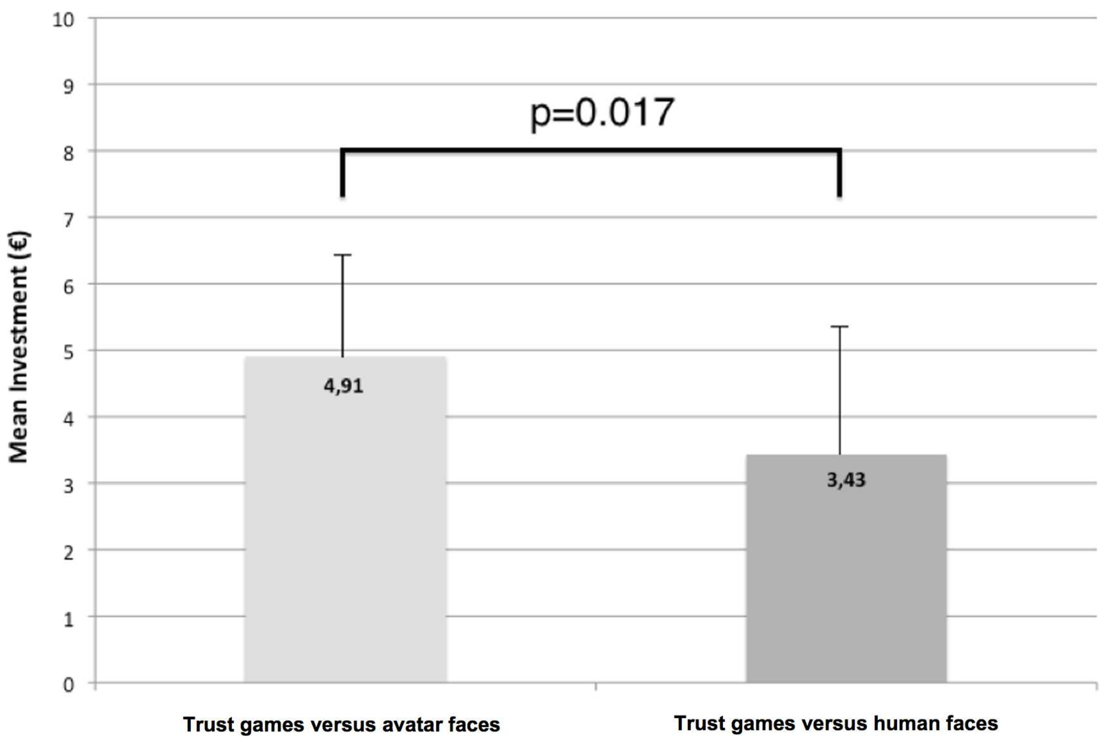


Fig 5. Results of overall mean investments in games versus avatar faces (left) and human faces (right) of the PD patients group.

doi:10.1371/journal.pone.0165998.g005

H2 states that PD patients show similar initial trust in avatar faces if compared to healthy controls. H2 is also supported by our data. There was no significant difference in the mean investment amount between the PD patients group and the control group in trust games played against avatar faces (mean patients: €4.91, SD 1.81; mean controls: €4.94, SD 1.55; two-sided unpaired t-test; $t = -0.07$; $p = 0.949$; see Fig 6). Further analyses to determine the effect size of this condition revealed a Cohen's d = 0.02 and an effect size r = 0.01.

PD patients, if compared to healthy controls, invested significantly less in games against human faces (mean PD patients: €3.43, SD 1.996; mean controls: €5.53, SD 1.56; two-sided unpaired t-test, $t = -3.70$; $p = 0.001$). The Game-of-Dice Task showed a significant difference in risk behavior between the patient group and the control group, with a larger number of high-risk choices (out of 18 choices) in the PD patient group (mean PD patients: 10.2, SD 3.928; mean controls: 7.05, SD 3.620; two-sided unpaired t-test; $t = 2.637$, $p = 0.012$).

Discussion

The results of our study confirm H1 that PD patients have significantly higher initial trust towards avatar faces when compared with human faces. Our data further support H2, which states that there is no significant difference between initial trust levels towards avatar faces between PD patients and healthy controls. Although it cannot be ruled out that a larger sample size could lead to a significant difference in this condition, we performed additional statistical analyses (Cohen's d), which indicates that there is indeed a negligible difference between mean investments of PD patients and healthy controls (see Fig 5). This is an intriguing result, because PD patients are known to invest less in a trust game, if compared to healthy controls, when playing against human faces [7]. Our control condition (Game-of-Dice task) confirms that

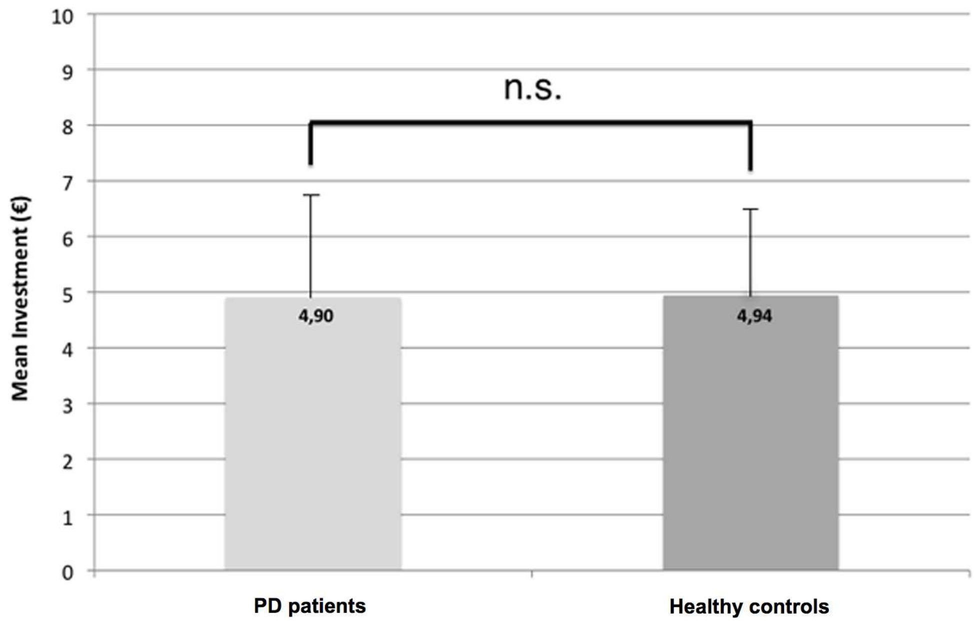


Fig 6. Results of overall mean investments in games versus avatar faces of PD patients (left) and healthy controls (right).

doi:10.1371/journal.pone.0165998.g006

lower trust in human faces is not a consequence of higher risk aversion in PD. Because of our study design (emotionless human faces and avatar faces, as well as no difference in colors used in the stimulus material) we can rule out effects of emotion recognition deficits (e.g. [88]) or color discrimination problems [89], known to affect PD patients, on our data.

Even though the results of our behavioral study do not make it possible to determine the underlying neurological mechanisms of the observed trust differences, our results have important implications for HCI in general and for specific HCI technologies used in medicine. Because evidence shows that the belief of interacting with either an avatar (user-controlled virtual character) or an agent (machine-controlled virtual character) does not result in significant differences with regard to the evaluation of the virtual character nor in different behavioral reactions [90], our discussion of implications is relevant for interaction of PD patients with both avatars and agents.

The findings of the present study are relevant to both computer scientists and programmers with a focus on HCI research, as well as medical scientists and physicians. Therefore, we discuss our results from a multidisciplinary perspective. We start the following discussion with HCI, where we group the implications in e-commerce, e-learning, and social media. A reflection on our results from the perspective of computer use in medicine follows, grouped into diagnosis and therapy in PD and patient-physician communication. Following the implications sections, we describe limitations of our study and close the paper with final remarks.

Implications for HCI

Avatars are used in several domains of HCI, such as e-commerce, e-learning, and social media.

In e-commerce, avatars can lead to a higher satisfaction with online retailers [91], because they can substitute a real-life salesperson (in the sense of an automated online assistant) and thereby reduce information overload [92], provide recommendations on suitable products [93], and enhance consumer decision-making (e.g. [94]). Furthermore, avatars have a significant effect on how consumer reviews are perceived by consumers [95]. This is important, because consumer reviews have a major impact on purchase decisions (e.g. [96]). Our data implies that online shops frequently used by PD patients could increase trust by using avatars as virtual sales consultants. It has been argued by Javor et al. [60] that all research involving patients or disabled persons and possibly affecting marketing should be evaluated from an ethical point of view, as disease specific weaknesses of these populations can be subject to target marketing. Our study might raise concerns whether PD patients can be targeted by online marketing measures using avatars. Importantly, the results of our study show that avatars are trusted to a similar extent by PD patients and healthy controls. Hence, the interaction of PD patients and avatars in a trust-related context such as online shopping does not seem to be affected by the disease process and therefore, in our understanding, the use of avatar salespersons in the interaction with PD patients is ethically uncritical.

E-learning is another application domain of avatars. In recent years, this domain has grown considerably, and the market for learning management systems has also exceeded growth predictions, currently worth around $2.55 billion worldwide according to the E-Learning Market Trends & Forecast 2014–2016. The use of virtual environments and avatars in e-learning systems play a key role in creating a sense of presence for users in the shared environment [97]. Grujic et al. [98] argue for the use of a virtual tutor in an electronic educational system. The tutor's major aim is not necessarily to teach students, but to emotionally respond to their actions with gestures and facial expressions. In a randomized controlled trial to evaluate the effectiveness of a patient education and health promotion program in the treatment of Parkinson's disease, the intervention group had significantly increased exercise, decreased "time off" (i.e., time periods of increasing symptoms mostly due to medication effects) and percentage of time off, reduced side effects, and decreased summary Parkinson's scores by approximately 10% [99]. There have been efforts to use e-learning for patient education in neurological patient populations, but these have not shown comparable results to face-to-face education (e.g. [100]). Given the positive effect of trust in the teacher on student performance [101], our data might help to improve patient education through e-learning by the use of avatars as tutors. Research in the field of Alzheimer's disease has already shown that patients suffering from neurological diseases might benefit from interaction with avatars [102]. Our results suggest that this finding could also hold true in the population of PD patients.

Avatars have also found their way into social media environments such as Facebook [103]. A recent review focusing on studies about the use of social media in patients and caregivers identified discussion forums as being highly popular (66.6%), and social networking sites (14.8%) and blogs/microblogs (14.1%) were the next most prevalent applications [104]. Several researchers highlight the importance of trust in social media environments (e.g. [105]). Dubois et al. [106] state that "[w]ith so much user interaction and content created, the question of whom and what to trust has become an increasingly important challenge on the web. A user is likely to encounter dozens if not hundreds of pieces of user-generated content each day, and some of it will need to be evaluated for trustworthiness. Trust information can help a user make decisions, sort and filter information, receive recommendations, and develop a context within a community with respect to whom to trust and why"(p.1).

A review of public health messages through social media indicates that more and more consumers turn to the internet for health related information and health organizations have therefore begun to turn to social media as a tool for connecting with the public, but only a very limited number of studies have analyzed the efficacy of social media in this context so far. Preliminary reports point to considerable outreach of social media applications and have the potential for engaging specific target audiences [107]. We argue for the use of avatars in public health social media messages aimed at PD patients in order to increase trust towards these contents. We could not identify a scientific study on the use of social media by PD patients. Hence, we make a call for research in this domain. It has already been discussed that avatars could be used to facilitate access to technologies for the elderly [108]. Our data support the idea of integrating avatars in the design of social media platforms to increase PD patients' trust, thereby increasing the inclusion of these patients in online social interaction.

Implications for the use of computers in medicine

In the medical field, several applications of avatars have emerged (e.g., diagnosis and therapy of neurological disorders and patient-physician communication).

Avatars have been used in the diagnosis and therapy of neurological disorders for almost 20 years [109]. An important part of therapy in PD is physical exercise that has been shown to positively affect physical functioning, health-related quality of life, strength, balance, and gait speed of PD patients [110]. Problems frequently occurring in intensive daily training are often motivational. Virtual environments have been used frequently in rehabilitation of PD in recent years to overcome these problems, because they contribute to motivating patients to exercise more [111]. In PD, two aspects of rehabilitation are of major concern. First, the above mentioned cardinal motor symptoms of the disease are being targeted. In this context, virtual environments often use avatars to represent the PD patient (e.g. [112]). In the light of evidence from cyberpsychology that players of online role-playing games sometimes identify themselves more strongly with the avatar compared to the real self (e.g. [113]), PD patient self-representation with an avatar seems to be an effective way to enhance rehabilitation. Second, behavioral and cognitive rehabilitation of non-motor symptoms (e.g. social interaction for behavioral symptoms or playing "Sudoku" for cognitive symptoms) [114] plays an important role, and avatars are often used in this context (e.g. [115]). Previous literature argues for the use of avatars in the treatment of mentalizing deficits of patients suffering from autism [116], and we argue that patients with PD might also benefit from such a training. Further, given the importance of trust in rehabilitation [117], our data supports the use of avatars in both motor and cognitive rehabilitation of PD patients.

With respect to avatars in patient-physician communication, a recent study advocates telemedicine in the care of PD patients and is arguing that"[t]ravel distance, growing disability, and uneven distribution of doctors limit access to care for most Parkinson's disease (PD) patients worldwide. Telemedicine, the use of telecommunications technology to deliver care at a distance, can help overcome these barriers"[118]. It has previously been argued that avatars could be used in patient communication to promote behavioral change, such as life style modification [119]. Our data show that avatar faces are trusted significantly more than human faces by PD patients, and this level of trust is similar to that of healthy controls. It can be theorized that communication with PD patients through avatars in the context of telemedicine could promote the patient-physician relationship and therefore improve health care in PD.

Limitations

The stimulus material consisted of unknown human faces and unknown avatar faces. It follows that results cannot be generalized to interactions with known counterparts. Furthermore, we make a call for studies replicating our investigation in order to substantiate the results of the present study.

As our study examined trust behavior of PD patients towards avatar faces (i.e., patients acted in the role of investor in our trust game), there is still missing evidence how PD patients would behave in the trustee role (i.e., the party who is given trust and subsequently has to reciprocate, or not). Thus, we make a call for future studies investigating PD patients' reciprocation of trust. Specifically, it will be rewarding to see whether differences exist in PD patients' interaction with both humans and avatars acting in the role of the investor. It might turn out that PD patients not only exhibit increased trust towards avatars when compared to humans; rather, it is possible that PD patients are also more trustworthy in their interactions with avatars than in their interaction with humans (i.e., they reciprocate more in the avatar condition than in the human condition in a trust game where they are given money by the investor).

Furthermore, we acknowledge that the results of the present study should be validated with a larger sample size. However, our statistically significant results based on a sample size of $n = 20$ in each experimental group along with a conservative level of significance ($p<0.05$) imply a large effect size. Additionally, the characteristics of our subjects (e.g. age, education level, all Caucasian, all Catholic) should be considered when interpreting our findings, because trust has been shown to correlate with those factors [69].

Finally, our study design using behavioral methodology does not allow for direct inferences on the exact nature or neurological basis of the differences in trust behavior of PD patients towards human faces and avatar faces. Therefore, we encourage studies using electrophysiological (e.g. electroencephalography, EEG) or functional brain imaging tools in order to shed light on the neural basis of trust behavior in PD.

Conclusion

Our study is a first step to examine human-avatar interaction at the nexus of behavioral neurology (neurological disorders affecting behavior) and PD. Our results suggest that avatars can be used to overcome behavioral deficits caused by a neurological disease, in our case PD patients' reduced trust in human social interaction. We hope that the present study instigates further interdisciplinary research in the domain of behavioral neurology and HCI. Advancements in this research field are urgently needed because neurological diseases and disorders are on the rise, and hence it is important to determine the positive and negative effects of IT use (e.g. in the form of avatars) for different patient populations.

Supporting Information

S1 File. File containing parts of the original data following recommendations from the local ethics committee.
(XLSX)

Author Contributions

Conceptualization: AJ GR WS RR.

Data curation: AJ RR.

Formal analysis: AJ RR.

Funding acquisition: AJ GR RR.

Investigation: AJ GR WS RR.

Methodology: AJ GR WS RR.

Project administration: AJ.

Resources: AJ GR WS.

Software: AJ.

Supervision: AJ RR.

Validation: AJ.

Visualization: AJ RR.

Writing – original draft: AJ RR.

Writing – review & editing: AJ GR WS RR.

References

1. Hirtz D, Thurman DJ, Gwinn-Hardy K, Mohamed M, Chaudhuri AR, Zalutsky R. How common are the "common" neurologic disorders? Neurology. 2007; 68(5): 326–337. doi: 10.1212/01.wnl.0000252807.38124.a3 PMID: 17261678
2. Hughes AJ, Ben-Shlomo Y, Daniel SE, Lees AJ. What features improve the accuracy of clinical diagnosis in Parkinson's disease: A clinicopathologic study. Neurology. 1992; 42(6): 1142–1142. PMID: 1603339
3. Aarsland D, Kurz MW. The epidemiology of dementia associated with Parkinson's disease. Brain Pathol. 2010; 20: 633–639. doi: 10.1111/j.1750-3639.2009.00369.x PMID: 20522088
4. Brand M, Labudda K, Kalbe E, Hilker R, Emmans D, Fuchs G, et al. Decision-making impairments in patients with Parkinson's disease. Behav Neurol. 2004; 15(3–4): 77–85. PMID: 15706051
5. Dujardin K, Blairy S, Defebvre L, Duhem S, Noël Y, Hess U, Destée A. Deficits in decoding emotional facial expressions in Parkinson's disease. Neuropsychologia. 2004; 42: 239–50. PMID: 14644109
6. Poletti M, Bonuccelli U. Personality traits in patients with Parkinson's disease: Assessment and clinical implications. J Neurol. 2012; 259(6): 1029–1038. doi: 10.1007/s00415-011-6302-8 PMID: 22083431
7. Javor A, Riedl R, Kirchmayr M, Reichenberger M, Ransmayr R. Trust behavior in Parkinson's disease: Results of a trust game experiment. BMC Neurol. 2015; 15: 126. doi: 10.1186/s12883-015-0374-5 PMID: 26227627
8. Frith CD, Frith U. The neural basis of mentalizing. Neuron. 2006; 50: 531–534. doi: 10.1016/j.neuron.2006.05.001 PMID: 16701204
9. Yu RL, Wu RM. Social brain dysfunctions in patients with Parkinson's disease: A review of theory of mind studies. Transl Neurodegener. 2013; 2(1): 7. doi: 10.1186/2047-9158-2-7 PMID: 23537376
10. Braak H, Bohl JR, Müller CM, Rüb U, de Vos RA, Del Tredici K. Stanley Fahn Lecture 2005: The staging procedure for the inclusion body pathology associated with sporadic Parkinson's disease reconsidered. Mov Disord. 2006; 21(12): 2042–2051. doi: 10.1002/mds.21065 PMID: 17078043
11. Schrag A, Jahanshahi M, Quinn N. What contributes to quality of life in patients with Parkinson's disease? J Neurol Neurosurg Psychiatry. 2000; 69(3):308–312. doi: 10.1136/jnnp.69.3.308 PMID: 10945804
12. Hartzband P, Groopman J. Untangling the Web-patients, doctors, and the Internet. N Engl J Med. 2010; 362(12): 1063–1066. doi: 10.1056/NEJMp0911938 PMID: 20335581
13. Pavlou PA, Liang H, Xue Y. Understanding and mitigating uncertainty in online exchange relationships: A principal agent perspective. Manag Inf Syst Q. 2007; 31: 1–32.
14. Begnum MEN. Challenges for Norwegian PC-Users with Parkinson's Disease–A Survey. In: Miesenberger K, Klaus J, Zagler W, Karshmer A, editors. Computers Helping People with Special Needs. Berlin, Heidelberg: Springer; 2010. pp. 292–299.
15. Biswas P, Robinson P. Modelling user interfaces for special needs. Accessible Design in the Digital World (ADDW); 2008.
16. Keates S, Trewin S. Effect of age and Parkinson's disease on cursor positioning using a mouse. In: Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility; 2005. pp. 68–75.
17. Cotten SR, Anderson WA, McCullough BM. Impact of internet use on loneliness and contact with others among older adults: Cross-sectional analysis. J Med Internet Res. 2013; 15(2): e39. doi: 10.2196/jmir.2306 PMID: 23448864
18. Karlsen KH, Tandberg E, Årsland D, Larsen JP. Health related quality of life in Parkinson's disease: a prospective longitudinal study. J Neurol Neurosurg Psychiatry. 2000; 69(5): 584–589. doi: 10.1136/jnnp.69.5.584 PMID: 11032608
19. Bailenson JN, Swinth K, Hoyt C, Persky S, Dimov A, Blascovich J. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. Presence (Camb). 2005; 14(4): 379–393.

20. Walther JB. Interpersonal effects in computer-mediated interaction a relational perspective. Commun Res. 1992; 19(1): 52–90.
21. Berger CR, Calabrese RJ. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. Hum Commun Res. 1975; 1(2): 99–112.
22. Schneider DJ, Hastorf AH, Ellsworth PC. Person perception. 2nd ed. Reading, MA: Addison-Wesley; 1979.
23. Kelley HH. The warm-cold variable in first impressions of persons. J Pers. 1950; 18(4): 431–439. PMID: 15428970
24. Hum NJ, Chamberlin PE, Hambright BL, Portwood AC, Schat AC, Bevan JL. A picture is worth a thousand words: A content analysis of Facebook profile photographs. Comput Human Behav. 2011; 27(5): 1828–1833.
25. MacDorman KF, Green RD, Ho CC, Koch CT. Too real for comfort? Uncanny responses to computer generated faces. Comput Human Behav. 2009; 25(3): 695–710. PMID: 25506126
26. Moser E, Derntl B, Robinson S, Fink B, Gur RC, Grammer K. Amygdala activation at 3T in response to human and avatar facial expressions of emotions. J Neurosci Methods. 2007; 161(1): 126–133. doi: 10.1016/j.jneumeth.2006.10.016 PMID: 17126910
27. Bailenson JN, Blascovich J, Beall AC, Loomis JM. Interpersonal distance in immersive virtual environments. Pers Soc Psychol Bull. 2003; 29: 819–833. doi: 10.1177/0146167203029007002 PMID: 15018671
28. Bailenson JN, Yee N, Merget D, Schroeder R. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. Presence (Camb). 2006; 15: 359–372.
29. Riedl R, Mohr PN, Kenning PH, Davis FD, Heekeren, HR. Trusting humans and avatars: A brain imaging study based on evolution theory. J Manage Inform Syst. 2014; 30(4): 83–114.
30. Donath J. Virtually trustworthy. Science. 2007; 317: 53–54. doi: 10.1126/science.1142770 PMID: 17615329
31. Park IH, Kim JJ, Jang HJ, Kim CH, Ku J, Kim IY, et al. Characteristics of social anxiety from virtual interpersonal interactions in patients with schizophrenia. Psychiatry. 2009; 72(1): 79–93. doi: 10.1521/psyc.2009.72.1.79 PMID: 19366296
32. Kim E, Jung YC, Ku J, Kim JJ, Lee H, Kim SY, et al. Reduced activation in the mirror neuron system during a virtual social cognition task in euthymic bipolar disorder. Prog Neuropsychopharmacol Biol Psychiatry. 2009; 33(8): 1409–1416. doi: 10.1016/j.pnpbp.2009.07.019 PMID: 19632283
33. Arias P, Robles-García V, Sanmartín G, Flores J, Cudeiro J. Virtual reality as a tool for evaluation of repetitive rhythmic movements in the elderly and Parkinson's disease patients. PloS One. 2012; 7(1): e30021. doi: 10.1371/journal.pone.0030021 PMID: 22279559
34. Cunningham LM, Nugent CD, Finlay DD, Moore G, Craig D. A review of assistive technologies for people with Parkinson's disease. Technol Health Care. 2009; 17(3): 269–279. doi: 10.3233/THC-2009-0547 PMID: 19641262
35. Yu W, Vuong C, Ingalls T. An interactive multimedia system for Parkinson's patient rehabilitation. In: Shumaker R, editor. Virtual and Mixed Reality-Systems and Applications. Heidelberg, Berlin: Springer; 2011. pp. 129–137.
36. Berg J, Dickhaut J, McCabe K. Trust, reciprocity, and social-history. Games Econ Behav. 1995; 10: 122–142.
37. McCabe KA, Smith V. A two person trust game played by naive and sophisticated subjects. Proc Natl Acad Sci U S A. 2000; 97: 3777–3781. doi: 10.1073/pnas.040577397 PMID: 10725349
38. McCabe K, Rigdon M, Smith V. Positive reciprocity and intentions in trust games. J Econ Behav Organ. 2003; 52: 267–275.
39. Zgaljardic DJ, Borod JC, Foldi NS, Mattis P. A review of the cognitive and behavioral sequelae of Parkinson's disease: relationship to frontostriatal circuitry. Cogn Behav Neurol. 2003; 16(4): 193–210. PMID: 14665819
40. Lewis SJ, Dove A, Robbins TW, Barker RA, Owen AM. Cognitive impairments in early Parkinson's disease are accompanied by reductions in activity in frontostriatal neural circuitry. J Neurosci. 2003; 23(15): 6351–6356. PMID: 12867520
41. Braak H, Braak E, Yilmazer D, de Vos RA, Jansen EN, Bohl J, et al. Amygdala pathology in Parkinson's disease. Acta Neuropathol. 1994; 88(6): 493–500. PMID: 7879596
42. Harding AJ, Stimson E, Henderson JM, Halliday GM. Clinical correlates of selective pathology in the amygdala of patients with Parkinson's disease. Brain 2002; 125(11): 2431–2445.

43. Adolphs R. Neural systems for recognizing emotion. Curr Opin Neurobiol. 2002; 12: 169–77. PMID: 12015233

44. Dujardin K, Blairy S, Defebvre L, Duhem S, Noël Y, Hess U, et al. Deficits in decoding emotional facial expressions in Parkinson's disease. Neuropsychologia. 2004; 42(2): 239–250. PMID: 14644109

45. Wu T, Wang J, Wang C, Hallett M, Zang Y, Wu X, et al. Basal ganglia circuits changes in Parkinson's disease patients. Neurosci Lett. 2012; 524(1): 55–59. doi: 10.1016/j.neulet.2012.07.012 PMID: 22813979

46. van der Vegt JP, Hulme OJ, Zittel S, Madsen KH, Weiss MM, Buhmann C, et al. Attenuated neural response to gamble outcomes in drug-naive patients with Parkinson's disease. Brain. 2013; 136(4): 1192–1203.

47. Ernst M, Paulus MP. Neurobiology of decision-making: a selective review from a neurocognitive and clinical perspective. Biol Psychiatry. 2005; 58: 597–604. doi: 10.1016/j.biopsych.2005.06.004 PMID: 16095567

48. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275 (5306): 1593–1599. PMID: 9054347

49. Tinaz S, Courtney MG, Stern CE. Focal cortical and subcortical atrophy in early Parkinson's disease. Mov Disord. 2011; 26(3): 436–441. doi: 10.1002/mds.23453 PMID: 21462259

50. Bodden M, Dodel R, Kalbe E. Theory of mind in Parkinson's disease and related basal ganglia disorders: A systematic review. Mov Disord. 2010; 25: 13–27. doi: 10.1002/mds.22818 PMID: 19908307

51. Riedl R, Javor A. The biology of trust: Integrating evidence from genetics, endocrinology and functional brain imaging. J Neurosci Psychol Econ. 2012; 5(2): 63–91.

52. Zak PJ, Knack S. Trust and growth. Econ J. 2011; 111(470): 295–321.

53. Bromiley P, Cummings LL. Transactions costs in organizations with trust. Res Negot O. 1995; 5: 219–250.

54. Rousseau DM, Sitkin SB, Burt RS, Camerer C. Not so different after all: A cross-discipline view of trust. Acad Manage Rev. 1998; 23(3): 393–404.

55. Gambetta D. Can we trust trust. In: Gambetta D, editor. Trust: Making and breaking cooperative relations. Oxford: Blackwell; 1988. pp. 213–237.

56. Coleman JS. Foundations of social theory. Harvard: University Press; 1994.

57. Robert LP, Denis AR, Hung YTC. Individual swift trust and knowledge-based trust in face-to-face and virtual team members. J Manage Inform Syst. 2009; 26(2): 241–279.

58. Fehr E. On the economics and biology of trust. J Eur Econ Assoc. 2009; 7: 235–266.

59. Fehr E. Social preferences and the brain. In: Glimcher PW, Fehr E, editors. Neuroeconomics: Decision making and the brain. Academic Press; 2009. pp. 215–232.

60. Javor A, Koller M, Lee N, Chamberlain L, Ransmayr G. Neuromarketing and consumer neuroscience: Contributions to neurology. BMC Neurol. 2013; 13(1): 13.

61. Riedl R, Hubert M, Kenning P. Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. Manag Inf Syst Q. 2010; 34: 397–428.

62. Zak PJ, Kurzban R, Matzner WT. The neurobiology of trust. Ann N Y Acad Sci. 2004; 1032(1): 224–227.

63. Rolls ET. Limbic systems for emotion and for memory, but no single limbic system. Cortex. 2015; 62: 119–157. doi: 10.1016/j.cortex.2013.12.005 PMID: 24439664

64. Baumgartner T, Heinrichs M, Vonlanthen A, Fischbacher U, Fehr E. Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. Neuron. 2008; 58: 639–650. doi: 10.1016/j.neuron.2008.04.009 PMID: 18498743

65. King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR. Getting to know you: Reputation and trust in a two-person economic exchange. Science. 2005; 308(5718): 78–83. doi: 10.1126/science.1108062 PMID: 15802598

66. Winston JS, Strange BA, O'Doherty J, Dolan RJ. Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nat Neurosci. 2002; 5(3): 277–283. doi: 10.1038/nn816 PMID: 11850635

67. Koscik TR, Tranel D. The human amygdala is necessary for developing and expressing normal interpersonal trust. Neuropsychologia. 2011; 49: 602–611. doi: 10.1016/j.neuropsychologia.2010.09.023 PMID: 20920512

68. Arsalidou M, Duerden EG, Taylor MJ. The centre of the brain: Topographical model of motor, cognitive, affective, and somatosensory functions of the basal ganglia. Hum Brain Mapp. 2013; 34: 3031–3054. doi: 10.1002/hbm.22124 PMID: 22711692

69. Schultz W. Behavioral theories and the neurophysiology of reward. Annu Rev Psychol. 2006; 57: 87–115. doi: 10.1146/annurev.psych.56.091103.070229 PMID: 16318590

70. Delgado MR, Frank RH, Phelps EA. Perceptions of moral character modulate the neural systems of reward during the trust game. Nat Neurosci. 2005; 8: 1611–1618. doi: 10.1038/nn1575 PMID: 16222226

71. Williamson O. E. (1993). Calculativeness, trust, and economic organization. J Law Econ. 1993; 36: 453–486.

72. Krueger F, McCabe K, Moll J, Kriegeskorte N, Zahn R, Strenziok M et al. Neural correlates of trust. Proc Natl Acad Sci U S A. 2007; 104(50): 20084–20089. doi: 10.1073/pnas.0710103104 PMID: 18056800

73. McCabe K, Houser D, Ryan L, Smith V, Trouard T. A functional imaging study of cooperation in two-person reciprocal exchange. Proc Natl Acad Sci U S A. 2001; 98: 11832–11835. doi: 10.1073/pnas.211415698 PMID: 11562505

74. Bhatt M, Camerer CF. Self-referential thinking and equilibrium as states of mind in games: FMRI evidence. Games Econ Behav. 2005; 52: 424–459.

75. Gibb WR, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. J Neurol Neurosurg Psychiatry. 1988; 51(6): 745–752. PMID: 2841426

76. Hughes AJ, Daniel SE, Lees AJ. Improved accuracy of clinical diagnosis of Lewy body Parkinson's disease. Neurology. 2001; 57(8): 1497–1499. PMID: 11673599

77. Alesina A, La Ferrara E. Who trusts others? J Public Econ. 2002; 85: 207–234.

78. Löwe BP. Gesundheitsfragebogen für Patienten PHQ-D: Manual; Komplettversion und Kurzform. Pfizer GmbH; 2001. German.

79. Weintraub D, Hoops S, Shea JA, Lyons KE, Pahwa R, Driver-Dunckley ED, et al. Validation of the questionnaire for impulsive-compulsive disorders in Parkinson's disease. Mov Disord. 2009; 24(10): 1461–1467. doi: 10.1002/mds.22571 PMID: 19452562

80. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State (a practical method for grading the state of patients for the clinician). J Psychiatr Res. 1975; 12: 189–198. PMID: 1202204

81. Sockeel P, Dujardin K, Devos D, Deneve C, Destée A, Defebvre L. The Lille apathy rating scale (LARS), a new instrument for detecting and quantifying apathy: Validation in Parkinson's disease. J Neurol Neurosurg Psychiatry. 2006; 77(5): 579–584. doi: 10.1136/jnnp.2005.075929 PMID: 16614016

82. Fahn S, Elton RL. Unified rating scale for Parkinson's disease. In: Recent developments in Parkinson's disease. Florham Park, New York: Macmillan; 1987. pp. 153–163.

83. Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stebbins GT. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. Mov Disord. 2007; 22(1): 41–47. doi: 10.1002/mds.21198 PMID: 17115387

84. Lundqvist D, Flykt A, Ohman A; 1998. Database: Karolinska Directed Emotional Faces [Internet]. Accessed: http://www.emotionlab.se/resources/kdef.

85. Oosterhof NN, Todorov A. The functional basis of face evaluation. Proc Natl Acad Sci U S A. 2008; 105: 11087–92. doi: 10.1073/pnas.0805664105 PMID: 18685089

86. Riedl R, Mohr P, Kenning P, Davis FD, Heekeren H. Trusting humans and avatars: behavioral and neural evidence. In: Proceedings of the 32nd international conference on information systems; 2011. pp. 1–23.

87. Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E. Oxytocin increases trust in humans. Nature. 2005; 435(7042): 673–676. doi: 10.1038/nature03701 PMID: 15931222

88. Saenz A, Doé de Maindreville A, Henry A, Labbey S, Bakchine S, Ehrlé N. Recognition of facial and musical emotions in Parkinson's disease. Eur J Neurol. 2013; 20(3): 571–577. doi: 10.1111/ene.12040 PMID: 23279689

89. Bertrand JA, Bedetti C, Postuma RB, Monchi O, Génier Marchand D, Jubault T, et al. Color discrimination deficits in Parkinson's disease are related to cognitive impairment and white-matter alterations. Mov Disord. 2012; 27(14): 1781–1788. doi: 10.1002/mds.25272 PMID: 23147270

90. Von der Puetten AM, Krämer NC, Gratch J, Kang SH. "It doesn't matter what you are!" Explaining social effects of agents and avatars. Comput Human Behav. 2010; 26(6): 1641–1650.

91. Holzwarth M, Janiszewski C, Neumann MM. The influence of avatars on online consumer shopping behavior. J Mark. 2006; 70(4): 19–36.

92. Maes P. Agents that reduce work and information overload. Commun ACM. 1994; 37(7): 30–40.

93. Rust RT, Kannan PK. E-service: A new paradigm for business in the electronic environment. Commun ACM. 2003; 46(6): 36–42.

94. Grenci RT, Todd PA. Solutions-driven marketing. Commun ACM. 2002; 45(3): 64–71.

95. Lee M, Kim M, Peng W. Consumer reviews: reviewer avatar facial expression and review valence. Internet Res. 2013; 23(2): 116–132.

96. Chevalier JA, Mayzlin D. The effect of word of mouth on sales: online book reviews. J Mark Res. 2006; 43(3): 345–54.

97. Monahan T, McArdle G, Bertolotto M. Virtual reality for collaborative e-learning. Comput Educ. 2008; 50(4): 1339–1353.

98. Grujic Z, Kovacic B, Pandzic IS. Building Victor-A virtual affective tutor. In: 10th International Conference on Telecommunications; 2009. pp. 185–189.

99. Montgomery EB Jr, Lieberman A, Singh G, Fries JF. Patient education and health promotion can be effective in Parkinson's disease: a randomized controlled trial. Am J Med. 1994; 97(5): 429–435. PMID: 7977431

100. Maasland E, Koudstaal PJ, Habbema JDF, Dippel DWJ. Effects of an individualized multimedia computer program for health education in patients with a recent minor stroke or transient ischemic attack-a randomized controlled trial. Acta Neurol Scand. 2007; 115(1): 41–48. doi: 10.1111/j.1600-0404.2006.00722.x PMID: 17156264

101. Lee SJ. The relations between the student-teacher trust relationship and school success in the case of Korean middle schools. Educ Stud. 2007; 33(2): 209–216.

102. Carrasco E, Epelde G, Moreno A, Ortiz A, Garcia I, Buiza C, et al. Natural interaction between avatars and persons with Alzheimer's disease. Berlin Heidelberg: Springer; 2008. pp. 38–45.

103. Vasalou A, Joinson A, Bänziger T, Goldie P, Pitt J. Avatars in social media: Balancing accuracy, playfulness and embodied messages. Int J Hum Comput Stud. 2008; 66(11): 801–811.

104. Hamm MP, Chisholm A, Shulhan J, Milne A, Scott SD, Given LM, et al. Social media use among patients and caregivers: a scoping review. BMJ Open. 2013; 3(5).

105. Kopton I, Sommer J, Winkelmann A, Riedl R, Kenning P. Users' trust building processes during their initial connecting behavior in social networks: Behavioral and neural evidence. In: Proceedings of the International Conference on Information Systems (ICIS); 2013. pp.1-12.

106. Dubois T, Golbeck J, Srinivasan A. Predicting trust and distrust in social networks. In: Privacy, security, risk and trust (passat) and 2011 IEEE Third International Conference on Social Computing (Socialcom); 2011. pp. 418–424.

107. Newbold KB, Campos S. Media and social media in public health messages: A systematic review. Hamilton: McMaster Institute of Environment and Health; 2011.

108. Noy D, Ribeiro P, Iurgel IA. Embodied virtual agents as a means to foster e-inclusion of older people. In: Biswas P, Duarte C, Langdon P, Almeida L, editors. A multimodal end-2-end approach to accessible computing. London: Springer; 2013. p. 135–154.

109. Rizzo AA, Buckwalter JG. The status of virtual reality for the cognitive rehabilitation of persons with neurological disorders and acquired brain injury. Stud Health Technol Inform. 1997; 39: 22–33. PMID: 10168917

110. Goodwin VA, Richards SH, Taylor RS, Taylor AH, Campbell JL. The effectiveness of exercise interventions for people with Parkinson's disease: A systematic review and meta-analysis. Mov Disord. 2008; 23(5): 631–640. doi: 10.1002/mds.21922 PMID: 18181210

111. Fong KN, Chow KY, Chan BC, Lam KC, Lee JC, Li TH, et al. Research Usability of a virtual reality environment simulating an automated teller machine for assessing and training persons with acquired brain injury. J Neuroeng Rehabil. 2010; 30: 7–19.

112. Mendes FADS, Pompeu JE, Lobo AM, da Silva KG, Oliveira TDP, Zomignani AP, et al. Motor learning, retention and transfer after virtual-reality-based training in Parkinson's disease–effect of motor and cognitive demands of games: A longitudinal, controlled clinical study. Physiotherapy. 2012; 98 (3): 217–223. doi: 10.1016/j.physio.2012.06.001 PMID: 22898578

113. Bessière K, Seay AF, Kiesler S. The ideal elf: identity exploration in World of Warcraft. Cyberpsychol Behav. 2007; 10: 530–535. doi: 10.1089/cpb.2007.9994 PMID: 17711361

114. Nombela C, Bustillo PJ, Castell PF, Sanchez L, Medina V, Herrero MT. Cognitive rehabilitation in Parkinson's disease: evidence from neuroimaging. Front Neurol. 2011; 2: 82. doi: 10.3389/fneur.2011.00082 PMID: 22203816

115. Klinger E, Chemin I, Lebreton S, Marié RM. Virtual action planning in Parkinson's Disease: A control study. Cyberpsychol Behav. 2006; 9(3): 342–347. doi: 10.1089/cpb.2006.9.342 PMID: 16780402

116. Fabri M, Elzouki SYA, Moore D. Emotionally expressive avatars for chatting, learning and therapeutic intervention. In: Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments. Berlin Heidelberg: Springer; 2007. pp. 275–285.

117. Manderson L, Warren N. The art or (re)learning to walk: Trust on the rehabilitation ward. Qual Health Res. 2010; 20(10): 1418–1432. doi: 10.1177/1049732310373105 PMID: 20555012

118. Achey M, Aldred JL, Aljehani N, Bloem BR, Biglan KM, Chan P, et al. The past, present, and future of telemedicine for Parkinson's disease. Mov Disord. 2014; 29: 871–883. doi: 10.1002/mds.25903 PMID: 24838316

119. Ruiz JG, Andrade AD, Roos BA. Medical avatars-an innovative approach to fostering health promotion and lifestyle change. Fed Pract. 2012; 29(4): 25–31.

Journal of Neuroscience, Psychology, and Economics
2017, Vol. 10, No. 1, 1–8

1937-321X/17/$12.00 http://dx.doi.org/10.1037/npe0000070

Oxytocin, Trust, and Trustworthiness: The Moderating Role of Music

René Riedl
University of Applied Sciences Upper Austria and University of Linz

Andrija Javor
University of Linz

David Gefen
Drexel University

Andrea Felten and Martin Reuter
University of Bonn

Evidence has indicated that the neuroactive hormone oxytocin is essential for prosocial behavior, particularly trust. Exogenous administration of oxytocin has been shown to increase trust in humans. However, one may argue that, except the administration of oxytocin in nonhealthy patient groups (e.g., those with autism or anxiety disorders) to alleviate negative symptoms, external administration of oxytocin has little relevance in normal life. Music, a ubiquitous stimulus in human society, has been shown to increase oxytocin in medical therapy scenarios. Considering this evidence, we conducted a trust game experiment with a sample of healthy humans and investigated music's effects on the (a) trustor's oxytocin levels (blood sample measurement), (b) investment amount (trust behavior measurement), and (c) perception of the other player's trustworthiness (self-report). The results of our exploratory study show that an increase in oxytocin levels over 40 trials in a trust game increased perceived trustworthiness in the no-music condition but had no impact on investment amount (i.e., trust behavior). Moreover, music had no effect on oxytocin, trust behavior, or perceived trustworthiness. Thus, unlike prior research showing that music listening may increase self-reported trust in another individual, in the present study we found no effect of music on trust (on either a physiological or behavioral level). We surmise that this finding is a result of both the type of music played during task execution and music preferences. Thus, future research must carefully manipulate music features (e.g., pitch, rhythm, timbre, tempo, meter, contour, loudness, and spatial location) and consider a listener's music preferences to better understand music's effects on physiological, perceived, and behavioral trust.

Keywords: hormone, oxytocin, music, trust, trust game

There has been mounting evidence that besides parturition and lactation, the neurohypophysis hormone oxytocin is crucial for prosocial behavior, including maternal attachment and pair bonding. Hence, oxytocin has been linked to approach behavior in both animals and humans (Donaldson & Young, 2008). Consistent with this finding, evidence has suggested that oxytocin is related to human trust behavior. Experimental research (Zak, Kurzban, & Matzner, 2004; Zak, Kurzban, & Matzner, 2005) found that, based on the trust game, oxytocin levels were significantly elevated in participants (trustees) who received a signal of trust (operationalized via the amount of money sent by the trustor) relative to a random monetary transfer; moreover, that research found that trustees who received an intentional trust signal returned significantly more to the trustor if compared to a random monetary transfer. Those findings suggest that perception of a trust signal may increase *endogenous* oxytocin levels, and this may result in reciprocation of trust (i.e., a trustee turns out to be trustworthy). In another trust game

This article was published Online First April 3, 2017.
René Riedl, Digital Business Management, School of Management, University of Applied Sciences Upper Austria and Business Informatics – Information Engineering, University of Linz; Andrija Javor, Kepler University Clinic, University of Linz; David Gefen, Decision Sciences and MIS, LeBow College of Business, Drexel University; Andrea Felten, Department of Psychology, University of Bonn; Martin Reuter, Department of Psychology and Center for Economics & Neuroscience, University of Bonn.
Special thanks for the constructive comments that were provided during the review process. This study was supported by Linzer Hochschulfonds Grant VII-508-2012/Ha.
Correspondence concerning this article should be addressed to René Riedl, Digital Business Management, School of Management, University of Applied Sciences, Upper Austria, Wehrgrabengasse 1-3, 4400 Steyr, Austria. E-mail: rene.riedl@fh-steyr.at

experiment, Kosfeld, Heinrichs, Zak, Fischbacher, and Fehr (2005) administered either oxytocin or a placebo to trustors via nasal spray. Oxytocin significantly increased the trustors' trust into the trustees. Thus, that study shows that *exogenously* administered oxytocin may increase trust in social interaction scenarios. It is important to note that oxytocin does not cross the blood–brain barrier, though there is evidence from nonhuman primate studies, as well as some weak evidence from human studies, that oxytocin in blood is positively correlated with oxytocin in the central nervous system (measured in cerebral spinal fluid; for details, see Dal Monte, Noble, Turchi, Cummins, & Averbeck, 2014).

In a more recent study, Ebert et al. (2013) replicated the Kosfeld et al. (2005) findings in healthy participants; however, contrary effects were found in borderline personality disorder patients. In the latter group, oxytocin had a trust-lowering effect that correlated with patients' history of childhood trauma. Baumgartner, Heinrichs, Vonlanthen, Fischbacher, and Fehr (2008), in another study, demonstrated that repeated betrayal of trust did not deteriorate trust behavior after intranasal oxytocin administration, whereas trust decreased in the placebo condition. With respect to sex differences, Yao et al. (2014) reported that oxytocin makes female, but not male, individuals less forgiving following betrayal of trust. Consistent with those findings, results of a brain-imaging study with trust-varying stimuli showed significant activation differences between men and women in trust-related brain regions that are dense in oxytocin receptors, such as striatal areas (Riedl, Hubert, & Kenning, 2010). A more complete analysis of oxytocin studies in the trust domain can be found in a recent review article (Riedl & Javor, 2012). Of importance, despite the presented evidence that oxytocin is related to human trust behavior, it has been argued that this relationship is not yet well understood. In a recent critical review of empirical research, Nave, Camerer, and McCullough (2015) concluded that "the cumulative evidence does not provide robust convergent evidence that human trust is reliably associated with OT [oxytocin] (or caused by it)" (p. 772).

Even though the relationship between oxytocin and human trust behavior is still an open question, one may argue that a pharmacological challenge like intranasal oxytocin intake constitutes a profound intervention in the body's equilibrium. Indeed, oxytocin treatment has been suggested to have implications for at least patients suffering from disturbed social behavior such as autism (e.g., Andari et al., 2010) or social anxiety disorder (e.g., Labuschagne et al., 2010). It is important to note that in the normal nonpsychopathological range of behavior it has been shown that trust behavior exhibits great variability and that that variability may also be related to differences in endogenous oxytocin levels (Riedl & Javor, 2012). This leads to the research question of this study, namely whether environmental stimuli can trigger trust behavior by increasing endogenous oxytocin levels. This study tests that through music. Music is an omnipresent stimulus in human society. It is well known that companies influence consumer behavior by music and that this influence is typically mediated by mood (for a review, see Garlin & Owen, 2006). It is not unlikely that music's effect on outcome variables such as shopping behavior is mediated by not only mood but also other factors, particularly trust perceptions. However, whether this is the case or not is currently unknown.

Against that background of previous research, the question arises whether music affects endogenous oxytocin release in a listener, which, in turn, may affect trust behavior. To the best of our knowledge, no other study has investigated this question. The only study that touches on this specific topic was carried out with a sample of convalescent people (Nilsson, 2009).[1] Specifically, that study evaluated the effect of bed rest with music on relaxation among patients who had just undergone heart surgery. A total of 40 patients who were at the time undergoing "open coronary artery bypass grafting and/or aortic valve replacement surgery" (Nilsson, 2009, p. 2153) were randomly allocated to one of two experimental conditions: music listening during bed rest ($n = 20$) or bed rest only ($n = 20$; the music was "soft, relaxing,

[1] In this context, we also mention a study conducted by Grape, Sandgren, Hansson, Ericson, and Theorell (2002). This study "explored the possible beneficial effects of singing on well-being" (p. 65). Although singing and listening to music (as investigated in our study) are different things, both are musical activities. Hence, it is interesting that Grape et al. reported that serum oxytocin concentrations increased significantly after a 45-min singing lesson (measurement took place before and 30 min after the lesson).

and included different melodies of 60 to 80 beats per minute [bpm] and was for 30 minutes with a volume of 50–60 dB," Nilsson, 2009, p. 2155). Relaxation was measured during bed rest the day after surgery by assessing multiple parameters, including plasma oxytocin and subjective relaxation levels, among others. The results indicated that in the music group oxytocin increased significantly in contrast to the control group and that self-reported relaxation also increased significantly more in the music group than in the control group. Thus, that study suggests that listening to music during bed rest after open-heart surgery may have positive effects on relaxation, on both a physiological (oxytocin) and a perceptual (self-reported) level.

However, it may be argued that a population that had just undergone surgery is likely to exhibit different physiological and psychological patterns in comparison to a normal and healthy population. Hence, in this exploratory study we investigated with a "discovery" sample whether, and if so how much, external music stimuli affect oxytocin levels in healthy participants in a trust game context and whether a potential physiological music effect also influences trust perceptions and trust behavior.

Method

This study was approved by the responsible local ethics committee in Upper Austria. All participants gave written informed consent. A licensed medical doctor (the second author of this article) performed intravenous blood sample collections, as well as supervised all experimental sessions.

Participants

Participants were recruited based on the university e-mail list and consisted of 30 right-handed male graduate- and doctoral-level students who had no history of neurological or psychiatric illnesses and who took no medication at the time the experiment was conducted. There was no significant age difference between the treatment (music: $n = 15$, mean age = 27.2, $SD = 5.3$) and control (no music: $n = 15$, mean age = 26.1, $SD = 6.2$) groups, $F(1, 28) = .256$, $p = .617$.

Procedure

Participants were told that the experiment consists of an economic game during which blood samples would be taken. We avoided the term *trust game* (Berg, Dickhaut, & McCabe, 1995) to rule out potential physiological and behavioral effects. Specifically, the participants were told that (a) both players would receive an initial endowment of €10, (b) they would play only the first move of the game (i.e., in the role of the trustor), (c) the amount they give to the other player would be multiplied by six so that the other player would receive six times what they were giving this other player, and (d) the other player could then reciprocate by paying back any amount. For example, if a participant invested €10 (this would be the total initial endowment), then the other player would receive €60, of which this other player could then send back any amount in between €0 and €70 to the participant (€70 = €10 [initial endowment] + €60 [received amount]). Each participant also rated post hoc the trustworthiness of each stimulus (faces and nicknames; explained next) on a 7-point-Likert scale.

In our version of the trust game, the second move was actually not played, to rule out effects such as learning or anger (due to eventual unfair reciprocation). Thus, participants had to make only the initial investment decision, that is, the first move in the game. As a consequence, we investigated initial trust, and not knowledge-based trust, in this study. Initial trust is the trust someone has in a stranger before having any interaction experience with the person. Knowledge-based trust, in contrast, is trust that emerges through engagement with the other person. The participants' playing partners were presented on a computer screen in the form of (a) face pictures (simulating real-world face-to-face interaction) and (b) male nicknames without any further information (simulating interaction in anonymous Internet environments).

Participants were informed that their compensation for the experiment would be the outcome of one randomly selected round out of all 40 rounds played during the experiment, but in fact, all participants received the same amount (€35). The faces for the experiment were taken from an existing database for which face trustworthiness has been prerated (Lundqvist, Flykt, & Öhman, 1998; Oosterhof & Todorov, 2008).

Nicknames were created for this study based on existing male Internet chat room nicknames. The protocol consisted of both 10 trustworthy and 10 untrustworthy faces and the same number of trustworthy and untrustworthy nicknames. Each participant played 40 rounds during the experiment, of which 20 were against "faces" and 20 against "nicknames." Faces and nicknames were presented in blocks of 20 rounds. The experimental protocol is illustrated in Figure 1.

After briefing the 30 participants, we randomly assigned them to the two experimental groups, with an equal number of participants in each group. Next, a catheter to collect multiple blood samples (BSs) was inserted into the arm of the participant. The first blood sample, BS_1, was taken 10 min after insertion of the catheter. In the treatment group, music was played in the background while the participant played the trust game. The music playing, which started after BS_1 had been taken, was Schubert's "Marche Militaire" in D major (stimulus administration: via loudspeakers at a volume of 50–60 dB). In the control group, no music was played. Six blood samples (BS_1, BS_2, BS_3, BS_4, BS_5, BS_6) were taken during the experiment.

Blood samples were taken before the first round of the game (BS_1), after every 10 rounds of the game (BS_2, BS_3, BS_4, BS_5), and 10 min after the end of the trust game (BS_6). After playing all rounds of the trust game, the participants were asked (a) their age and whether they (b) recognized the music if they were in the treatment group, (c) liked the music, and (d) whether they knew the musical piece before the experiment (i.e., familiarity). Additionally, (e) disposition to trust was measured based on an existing instrument (Gefen, 2000). Also, (f) each participant rated post hoc the trustworthiness of all faces and nicknames on a 7-point-Likert scale.

Hormone Analyses and Sample Collection

Blood samples were collected through the catheter into prechilled ethylenediaminetetraacetic acid tubes and kept cold during every step of processing. Blood samples were immediately cold-centrifuged at 1,600 g for 15 min at 4 °C. Plasma was then transferred into chilled tubes and frozen at −80 °C until assay. Extraction of peptides from plasma as well as fluorescent immunoassay analyses of the samples were performed following the manufacturer protocol in the laboratory of the Department of Psychology, University of Bonn, using the oxytocin fluorescent immunoassay kit developed by the company Phoenix Pharmaceuticals (Burlingame, California). The inter- and intraassay co-

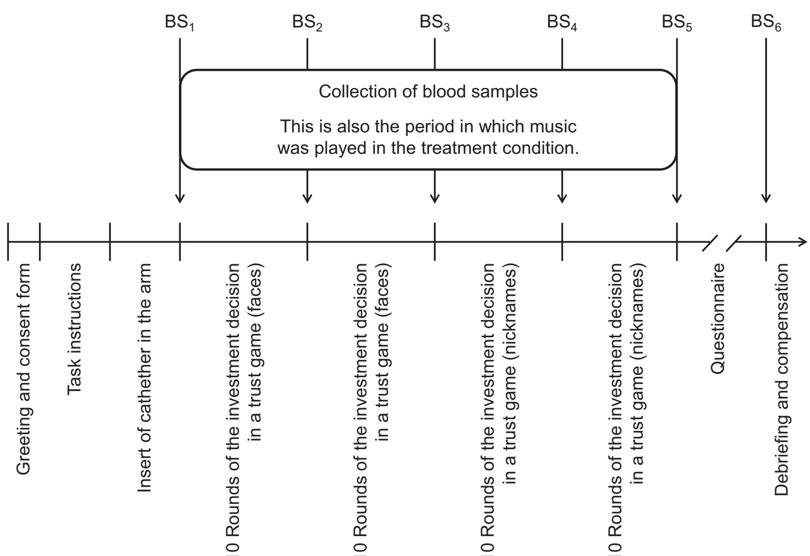


Figure 1. Experimental protocol. BS = blood sample. This protocol shows the experimental activities in sequential order (from left to right).

efficients of variability were <15% and <10%, respectively. All experimental sessions were conducted between 9.00 a.m. and 1.00 p.m.

Statistical Analyses

The effects of music on trusting behavior, the perceived trustworthiness of the other players, and oxytocin were analyzed by means of three separate repeated-measures analyses of variance (ANOVAs) with music (music or no music) as the between-subjects factor. In the respective analyses, trust behavior (four levels: Faces 1–10, Faces 11–20, Nicknames 1–10, Nicknames 11–20), perceived trustworthiness (four levels: Faces 1–10, Faces 11–20, Nicknames 1–10, Nicknames 11–20), and oxytocin levels (six time points of measurement) constituted the within-subject factor. Moreover, composite trust and trustworthiness scores were computed by summing up the responses of all 40 trials. These composite scores were tested for significant group difference (music or no music) by means of ANOVAs.

In an additional analysis, individual change of oxytocin over time was assessed by calculating a linear fit for each participant based on the person's respective six oxytocin values. The regression coefficients (beta weights [*b*s], slope of the regression line) provide information on the individual oxytocin-level changes across the experiment. Mean regression coefficients were compared between the experimental conditions (by means of ANOVA) and correlated with scores of trust and trustworthiness.

Before we report our results, we stress that the present study is exploratory in nature. Because of the relatively small sample size of 30 participants ($n = 15$ in the music condition and $n = 15$ in the no-music condition), replication studies are needed. This need for replications is independent of whether one reports statistically significant effects.

Results

Because it is known that oxytocin levels decrease during the day, showing higher levels in the morning and lower levels in the evening, two factors might counteract this natural oxytocin drop in the current experiment: the trust game and the experimental manipulation (i.e., listening to music).

On a descriptive level, results showed higher oxytocin levels across all six time points of measurement in the no-music condition when compared to the music condition (see Figure 2). However, the results of a repeated-measures analysis of variance yielded neither significant main effects of group, $F(1, 28) = 1.93$, $p = .176$, and time, $F(3.075, 86.090) = .692$, $p = .630$, nor a significant Time × Group interaction effect, $F(3.075, 86.090) = 1.74$, $p = .164$.[2] The insignificant statistical results may be the consequence of the relatively small sample consisting of only 15 participants each in the experimental and treatment groups. The effect size of the Time × Group interaction was a partial eta-square of 5.8%, suggesting a noticeable difference in measured oxytocin level change between the two experimental groups.

To quantify changes of oxytocin over time, we calculated a linear fit for each participant. The regression coefficients (beta weights [*b*s], slope of the regression line) provide information on the oxytocin level changes during the experiment. As expected based on the results of the repeated-measures analysis of variance, the slopes in the two experimental conditions did not differ significantly, $F(1, 29) = 1.13$, $p = .298$, although an increase in oxytocin over time was observable in the music condition ($b = .074$), whereas oxytocin levels decrease in the no-music condition ($b = -1.280$; see Figure 2).

However, we found significant correlations between the oxytocin slope and measures of trustworthiness but only in the no-music condition. Here the oxytocin slope was significantly correlated with trustworthiness of faces ($r = .550$, $p = .023$), nicknames ($r = .580$, $p = .023$), and the trustworthiness total scale (con-

[2] Both groups showed no significant differences in measures of trust (behavioral measure: investment amount in the trust game) and trustworthiness (7-point-Likert scale): trust faces ($M_{\text{no music}} = 86.067$, $SD = 28.273$; $M_{\text{music}} = 98.400$, $SD = 26.216$), $F(1, 28) = 1.54$, $p = .226$; trust nicknames ($M_{\text{no music}} = 108.267$, $SD = 36.916$; $M_{\text{music}} = 128.000$, $SD = 30.308$), $F(1, 28) = 2.56$, $p = .121$; trust total ($M_{\text{no music}} = 194.333$, $SD = 60.636$; $M_{\text{music}} = 226.400$, $SD = 49.838$), $F(1, 28) = 2.50$, $p = .125$; trustworthiness faces ($M_{\text{no music}} = 76.733$, $SD = 13.150$; $M_{\text{music}} = 80.133$, $SD = 12.794$), $F(1, 28) = 0.52$, $p = .479$; trustworthiness nicknames ($M_{\text{no music}} = 91.867$, $SD = 17.999$; $M_{\text{music}} = 102.600$, $SD = 13.378$), $F(1, 28) = 3.44$, $p = .074$; trustworthiness total ($M_{\text{no music}} = 168.600$, $SD = 27.959$; $M_{\text{music}} = 182.733$, $SD = 20.225$), $F(1, 28) = 2.52$, $p = .124$.

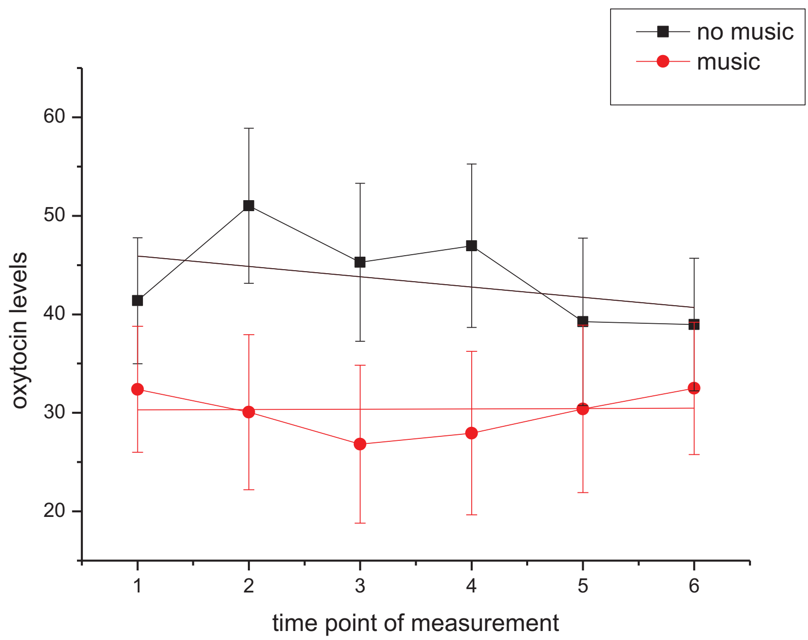


Figure 2. Oxytocin levels (pg/ml) in the treatment (music) and control (no music) groups across the six time points of measurement. Error bars indicate standard error of the means. See the online article for the color version of this figure.

taining faces and nicknames; r = .632, p = .011). There were consistently positive correlations between the oxytocin slope and trust in the no-music condition, although these correlations were not statistically significant. It is interesting that the correlations between oxytocin slope and measures of trust and trustworthiness were all negative, but not significant, in the music treatment condition.

Discussion

Based on previous studies that demonstrated the effects of oxytocin on trust and the trust elevating effects of music, this exploratory study used a trust game based on healthy participants to investigate the effects of music versus no music on endogenous oxytocin levels, trust behavior, and perceived trustworthiness of the other players. Despite the absence of a significant effect of music on oxytocin levels, trust, or trustworthiness in a between-subjects design, individual changes in oxytocin levels during the trust game were related to trustworthiness ratings. However, this effect was observable in only the no-music condition. Thus, music had no significant effect on oxytocin (yet, the effect size was not negligible), behavioral trust, and perceived trustworthiness. This result may seem inconsistent with results of prior research that showed that listening to music may increase self-reported trust in another individual (Anshel & Kipper, 1988).

However, because the *type of music* used in our study and in the prior study (Anshel & Kipper, 1988) differed significantly (Schubert's "Marche Militaire" in D major in the present study [i.e., classical music with no lyrics] vs. "popular Israeli songs," p. 149), this stimulus difference could have caused the difference in results. We do not know what songs exactly were played in the Anshel and Kipper (1988) study, but "Marche Militaire" is an invigorating martial music piece, potentially reducing oxytocin levels.[3] Our argument is based on evidence showing that music genre may strongly affect the consequences of music (e.g., Freeman, 2000). Music is characterized by many

[3] Possibly, the martial nature of the music, and the implied antisocial behavior of armies, may induce participants to be less social than they might normally be.

features, including pitch, rhythm, timbre, tempo, meter, contour, loudness, and spatial location (e.g., Levitin & Tirovolas, 2009). These features affect how music is perceived, as well as its effects. As an example, Gomez and Danuser (2007) demonstrated that perception of different musical features leads to significant differences in (a) physiological parameters such as respiration, skin conductance, and heart rate and (b) self-reported pleasantness and arousal.

Against the background of those conclusions, most likely musical features and their perception by the listener produce music's trust effects. Thus, future research with careful manipulation of music features is needed to better understand the effects of music on trust.

The positive association between oxytocin changes (i.e., increases across the experiment) and trustworthiness indicates that the neurohypophysis hormone is likely to promote feelings of trustworthiness when confronted with social interaction partners. Probably these prosocial feelings are blurred by the individual differences in the perception of the music being played (here classical music, which is rather old-fashioned to the young participants). Thus, future studies should consider *music preference.*[4] Of importance, previous research has shown that musical preference may significantly affect music's physiological effects. Specifically, a study by Vanderark and Ely (1993) revealed that musical preference significantly influenced listeners' cortisol response. Because cortisol and oxytocin have antagonistic effects (e.g., Ditzen et al., 2009; Heinrichs, Baumgartner, Kirschbaum, & Ehlert, 2003), careful consideration of a listener's music preference, in addition to manipulation of music features, is indispensable to develop a clear picture on music's effects on trust.

Conclusion

The direction and magnitude of the effects in this study suggest that the relationship between music, oxytocin, trusting behavior, and trustworthiness is complex. An increase in oxytocin levels over 40 trials in a trust game was associated with an increase in the perceived trustworthiness of interaction partners in the no-music condition but had no apparent impact on trust behavior. Moreover, music had no effect on oxytocin, trust behavior, or perceived trustworthiness. Future research should examine the influence of musical genres and other musical features, as well as how these are perceived by the participants, on oxytocin, trustworthiness perceptions, and behavioral trust.

[4] In the present study, we asked participants in the music condition ($n = 15$) after playing all rounds of the trust game whether they liked the music (yes: 13, no: 2). It follows that there is too little variance in data to determine possible preference effects on outcome variables. We recommend that future studies assess music preference in a pretest so that enough variance in data is likely to exist in the main study. Moreover, we recommend using an ordinal scale to measure music preference in future studies.

References

Andari, E., Duhamel, J.-R., Zalla, T., Herbrecht, E., Leboyer, M., & Sirigu, A. (2010). Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. *PNAS: Proceedings of the National Academy of Sciences of the United States of America, 107,* 4389–4394. http://dx.doi.org/10.1073/pnas.0910249107

Anshel, A., & Kipper, D. A. (1988). The influence of group singing on trust and cooperation. *Journal of Music Therapy, 3,* 140–155.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58,* 639–650. http://dx.doi.org/10.1016/j.neuron.2008.04.009

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10,* 122–142. http://dx.doi.org/10.1006/game.1995.1027

Dal Monte, O., Noble, P. L., Turchi, J., Cummins, A., & Averbeck, B. B. (2014). CSF and blood oxytocin concentration changes following intranasal delivery in macaque. *PLoS ONE 9*(8), e103677.

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological Psychiatry, 65,* 728–731.

Donaldson, Z. R., & Young, L. J. (2008, November 7). Oxytocin, vasopressin, and the neurogenetics of sociality. *Science, 322,* 900–904. http://dx.doi.org/10.1126/science.1158668

Ebert, A., Kolb, M., Heller, J., Edel, M. A., Roser, P., & Brüne, M. (2013). Modulation of interpersonal trust in borderline personality disorder by intranasal oxytocin and childhood trauma. *Social Neuroscience, 8,* 305–313. http://dx.doi.org/10.1080/17470919.2013.807301

Freeman, W. J. (2000). A neurobiological role of music in social bonding. In N. W. Merkur & S. Brown (Eds.), *The origins of music* (pp. 411–424). Cambridge, MA: MIT Press.

Garlin, F. V., & Owen, K. (2006). Setting the tone with the tune: A meta-analytic review of the effects of background music in retail settings. *Journal of Business Research, 59,* 755–764. http://dx.doi.org/10.1016/j.jbusres.2006.01.013

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega: The International Journal of Management Sciences, 28,* 725–737.

Gomez, P., & Danuser, B. (2007). Relationships between musical structure and psychophysiological measures of emotion. *Emotion, 7,* 377–387. http://dx.doi.org/10.1037/1528-3542.7.2.377

Grape, C., Sandgren, M., Hansson, L.-O., Ericson, M., & Theorell, T. (2002). Does singing promote well-being? An empirical study of professional and amateur singers during a singing lesson. *Integrative Physiological & Behavioral Science, 38,* 65–74. http://dx.doi.org/10.1007/BF02734261

Heinrichs, M., Baumgartner, T., Kirschbaum, C., & Ehlert, U. (2003). Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. *Biological Psychiatry, 54,* 1389–1398. http://dx.doi.org/10.1016/S0006-3223(03)00465-7

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005, June 2). Oxytocin increases trust in humans. *Nature, 435,* 673–676. http://dx.doi.org/10.1038/nature03701

Labuschagne, I., Phan, K. L., Wood, A., Angstadt, M., Chua, P., Heinrichs, M., . . . Nathan, P. J. (2010). Oxytocin attenuates amygdala reactivity to fear in generalized social anxiety disorder. *Neuropsychopharmacology, 35,* 2403–2413. http://dx.doi.org/10.1038/npp.2010.123

Levitin, D. J., & Tirovolas, A. K. (2009). Current advances in the cognitive neuroscience of music. *Annals of the New York Academy of Sciences, 1156,* 211–231. http://dx.doi.org/10.1111/j.1749-6632.2009.04417.x

Lundqvist, D., Flykt, A., & Öhman, A. (1998). Karolinska Directed Emotional Faces [Database of standardized facial images]. Stockholm, Sweden: Psychology Section, Department of Clinical Neuroscience, Karolinska Institute.

Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science, 10,* 772–789.

Nilsson, U. (2009). Soothing music can increase oxytocin levels during bed rest after open-heart surgery: A randomised control trial. *Journal of Clinical Nursing, 18,* 2153–2161. http://dx.doi.org/10.1111/j.1365-2702.2008.02718.x

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *PNAS: Proceedings of the National Academy of Sciences of the United States of America, 105,* 11087–11092. http://dx.doi.org/10.1073/pnas.0805664105

Riedl, R., Hubert, M., & Kenning, P. (2010). Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *Management Information Systems Quarterly, 34,* 397–428.

Riedl, R., & Javor, A. (2012). The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics, 5,* 63–91. http://dx.doi.org/10.1037/a0026318

VanderArk, S. D., & Ely, D. (1993). Cortisol, biochemical, and galvanic skin responses to music stimuli of different preference values by college students in biology and music. *Perceptual and Motor Skills, 77,* 227–234. http://dx.doi.org/10.2466/pms.1993.77.1.227

Yao, S., Zhao, W., Cheng, R., Geng, Y., Luo, L., & Kendrick, K. M. (2014). Oxytocin makes females, but not males, less forgiving following betrayal of trust. *International Journal of Neuropsychopharmacology, 17,* 1785–1792. http://dx.doi.org/10.1017/S146114571400090X

Zak, P. J., Kurzban, R., & Matzner, W. T. (2004). The neurobiology of trust. *Annals of the New York Academy of Sciences, 1032,* 224–227. http://dx.doi.org/10.1196/annals.1314.025

Zak, P. J., Kurzban, R., & Matzner, W. T. (2005). Oxytocin is associated with human trustworthiness. *Hormones and Behavior, 48,* 522–527. http://dx.doi.org/10.1016/j.yhbeh.2005.07.009

Received October 15, 2016
Revision received February 28, 2017
Accepted March 5, 2017 ■

Electronic Markets
https://doi.org/10.1007/s12525-022-00594-4

RESEARCH PAPER

Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions

René Riedl[1,2]

Received: 11 February 2022 / Accepted: 7 September 2022

Abstract
Artificial intelligence (AI) refers to technologies which support the execution of tasks normally requiring human intelligence (e.g., visual perception, speech recognition, or decision-making). Examples for AI systems are chatbots, robots, or autonomous vehicles, all of which have become an important phenomenon in the economy and society. Determining which AI system to trust and which not to trust is critical, because such systems carry out tasks autonomously and influence human-decision making. This growing importance of trust in AI systems has paralleled another trend: the increasing understanding that user personality is related to trust, thereby affecting the acceptance and adoption of AI systems. We developed a framework of user personality and trust in AI systems which distinguishes universal personality traits (e.g., Big Five), specific personality traits (e.g., propensity to trust), general behavioral tendencies (e.g., trust in a specific AI system), and specific behaviors (e.g., adherence to the recommendation of an AI system in a decision-making context). Based on this framework, we reviewed the scientific literature. We analyzed N = 58 empirical studies published in various scientific disciplines and developed a "big picture" view, revealing significant relationships between personality traits and trust in AI systems. However, our review also shows several unexplored research areas. In particular, it was found that prescriptive knowledge about how to design trustworthy AI systems as a function of user personality lags far behind descriptive knowledge about the use and trust effects of AI systems. Based on these findings, we discuss possible directions for future research, including adaptive systems as focus of future design science research.

Keywords Artificial Intelligence (AI) · Big Five traits · Machine learning (ML) · Personality · Review · Trust · Trust propensity

JEL classification D91 · M15 · O33

Introduction

McCarthy et al. (1955) referred to Artificial Intelligence (AI) as "making a machine behave in ways that would be called intelligent if a human were so behaving" (p. 11). In his seminal book on AI history, Nilsson (2010) specifies that intelligence is "that quality that enables an entity to function appropriately and with foresight in its environment [… and] because 'functioning appropriately and with foresight' requires so many different capabilities, depending on the environment, we actually have several continua of intelligences" (book preface). Thus, despite the fact that we observe an ongoing discussion on AI definitions in the scientific literature (that mainly results from the way how intelligence is conceptualized), in recent years AI systems such as chatbots, recommendation agents, autonomous vehicles,

This article is part of the Topical Collection on Trust in artificial intelligence

Responsible Editor: Wolfgang Maass

René Riedl
rene.riedl@fh-steyr.at; rene.riedl@jku.at

[1] Digital Business Institute, School of Business and Management, University of Applied Sciences Upper Austria, Wehrgrabengasse 1-3, 4400 Steyr, Austria

[2] Institute of Business Informatics – Information Engineering, Johannes Kepler University Linz, Altenberger Strasse 69, 4040 Linz, Austria

and robots have become a dominating topic, both in science and practice (e.g., Berente et al., 2021; Dwivedi et al., 2021). The success of integrating AI systems into society and the economy significantly depends on users' trust (e.g., Glikson & Woolley, 2020; Jacovi et al., 2021). AI systems increasingly carry out tasks autonomously and significantly influence human-decision making. Hence, determining which AI system to trust and which not to trust has become crucial. Trust in systems and machines which turn out not to be trustworthy may lead to undesired consequences (Lee & Moray, 1992). These may range from unfavorable purchases (e.g., when using recommendation agents in online shops) to death (e.g., when an autonomous vehicle causes a fatal accident). What follows is that the decision to trust AI systems always comes with some form of vulnerability. This explains why many novel AI systems are not always embraced by people, despite their enormous potential to make organizations more efficient and human life more comfortable (Collins et al., 2021). Considering this *enormous relevance of users' trust for the acceptance and adoption of AI systems*, it is no surprise that scientific research has studied this phenomenon extensively. In the past decade, several reviews on trust in autonomous systems and AI systems were published (Glikson & Woolley, 2020; Hoff & Bashir, 2015; Siau & Wang, 2018; Thiebes et al., 2021), and also meta-analyses are available (Hancock et al., 2011; Schaefer et al., 2016).

This growing importance of trust in AI systems has paralleled another trend: the *increasing relevance of personality research in the fields of AI, robotics, and autonomous systems* (Matthews et al., 2021). The American Psychological Association (APA) defines personality as "individual differences in characteristic patterns of thinking, feeling and behaving".[1] Other conceptualizations highlight the stability of the characteristics over time. For example, Mount et al. (2005) indicate that personality traits can be defined as "characteristics that are *stable over time* [… which] provide the reasons for the person's behavior […] they reflect who we are and in aggregate determine our affective, behavioral, and cognitive style" (pp. 448-449, italics added). In a more recent paper, Montag and Panksepp (2017) define personality as a set of "*stable* individual differences in cognitive, emotional and motivational aspects of mental states that result in stable behavioral action […] especially emotional tendencies" (p. 1, italics added).

Considering these definitions, it is obvious that human personality and trust tendencies must be related, a fact that also holds true in the specific context of trust in AI systems. However, to date, the role of personality in the context of trust in AI systems has *not* been investigated in the form of a systematic review of existing empirical evidence. Rather, existing works only briefly mention user personality as a factor related to trust in automation and AI systems (Glikson & Woolley, 2020; Hoff & Bashir, 2015; Siau & Wang, 2018; Thiebes et al., 2021). While this confirms the relevance of this article's topic, these brief references to user personality do not constitute systematic analysis of the existing knowledge. Moreover, recently Matthews et al. (2021) made the following statement in their paper on personality research in robotics, autonomous systems, and AI: "Person factors in trust in machines have been a relatively neglected aspect of research […]" (p. 3). User personality is a major person factor. In another recent paper, Jacovi et al. (2021) make an explicit call for personality research in the context of trust in AI systems—they write: "Personal Attributes of the Trustor […] Future work in this area may incorporate elements of the personal attributes of the trustor into the model, such as personality" (p. 633).

The fact that personality research in the context of trust in AI systems has not been examined systematically in the form of a review is problematic for at least four reasons.

First, as it has turned out after analyzing the extant literature, the existing empirical studies are published in various scientific disciplines, including computer science and robotics, ergonomics, human-computer interaction, information systems (IS), and psychology. Hence, a cumulative research tradition hardly exists. What follows is that to date a "*big-picture*" *view* of user personality and trust in the context of AI systems is not available. Thus, it is a fruitful scientific endeavor to *integrate empirical results across discipline boundaries and contexts*.

Second, investigation of the development of research at the nexus of user personality and trust in AI systems contributes to identity development in this important interdisciplinary and relatively new field. Based on a review of N = 58 empirical papers published in the period 2008-2021, in the current article we integrate available, but highly fragmented literature that is published in outlets across various scientific disciplines. Examination of a research field's development may provide valuable *insights into the future development*. In an essay on the identity of the IS discipline, Klein and Hirschheim (2008) write that "a shared sense of history provides the ultimate grounding and background information (preunderstanding) for communication in large and diverse collectives such as societies (and by extension to diverse disciplines)" (p. 298). Bearing this argument in mind, a main motivation of the present paper is to provide insights into the status of personality research in the context of trust in AI systems, thereby contributing to identity development. Knowing the history of the research field, even if it is short, facilitates identity formation by *identification of emergent thematic and methodological patterns*. In short, knowing

[1] https://www.apa.org/topics/personality (accessed on January 16, 2022).

one's past is important for coping with future challenges (Webster & Watson, 2002).

Third, personality has been identified as a major individual difference variable (along with gender, age, and culture) which affects trust in the context of automation and AI systems, and it has been argued that this fact has far-reaching *design consequences* (e.g., Hoff & Bashir, 2015). For example, Tapus et al.'s (2008) robot study demonstrated the feasibility of "a behavior adaptation system capable of adjusting its social interaction parameters (e.g., interaction distances/proxemics, speed, and vocal content) […] based on the user's personality traits and task performance" (p. 169). Importantly, a *sound understanding of users' personality and trust constitutes a foundation for the successful development of adaptive systems*. In general, such systems automatically adapt in real time based on users' states or traits to improve human-technology interactions (e.g., vom Brocke et al., 2020). Development of adaptive systems has become an important topic in IS research (e.g., Astor et al., 2013; Demazure et al., 2021) and also beyond IS, as signified by developments in disciplines like affective computing (e.g., Poria et al., 2017).

Fourth, a major motivation to study personality research in the context of trust in AI systems is that IS research has ignored this personality focus thus far. Maier (2012) and recently Sindermann et al. (2020) reviewed personality research in the IS discipline and report two thematic domains in which personality has played a major role: as antecedent in technology acceptance and adoption decisions, and as characteristic of computer personnel which influences further downstream variables (e.g., willingness to change job). Regarding acceptance and adoption, it is critical to emphasize that the extant IS literature had a focus on traditional applications, including enterprise systems and online shops without any AI component (e.g., Devaraj et al., 2008; McElroy et al., 2007). Therefore, while personality research has its place in the IS discipline, the present paper's specific thematic focus has not been addressed thus far.

The article is structured as follows. In Sect. 2, we develop a theoretical foundation for the present article. Specifically, we summarize foundations of personality research, as well as major insights into trust in the context of AI systems. In Sect. 3, we outline our review methodology. Sect. 4 presents major results of the review. Sect. 5 discusses the results, identifies several unexplored research areas, and outlines several fruitful directions for future research. In Sect. 6, we specifically discuss design implications and describe the development of adaptive systems as possible focus of future design science research. In Sect. 7, this paper's limitations and a concluding statement are provided. Altogether, the present article contributes to a better understanding of an increasingly important phenomenon in the digital economy and society, namely the influence of user personality on trust in AI systems, which, in turn, significantly affects the acceptance and adoption decisions. Therefore, the current topic is not only essential from an academic perspective, but also from a practice point of view.

Theoretical foundations

Personality

Human personality refers to stable characteristics that determine individual differences in thoughts, emotions, and behavior (Funder, 2001). Thus, personality affects how people react to all forms of stimuli, including situations in which humans interact with AI systems.

Today consensus exists that the *Five-factor Model*—in short, *Big Five*—constitutes the "broadest level of abstraction" to conceptualize human personality, where "each dimension summarizes a large number of distinct, more specific personality characteristics" (John & Srivastava, 1999, p. 105).[2] Initially described in the 1960s (Tupes & Christal, 1961), several research groups have contributed independently to the development of the model in the past decades (Cattell et al., 1970; Goldberg, 1990; McCrae & Costa, 1987). The model comprises the following five factors: extraversion ("an energetic approach toward the social and material world", example traits: sociability, activity, assertiveness, positive emotionality), agreeableness ("contrasts a prosocial and communal orientation toward others with antagonism", example traits: altruism, tender-mindedness, trust), conscientiousness ("describes socially prescribed impulse control that facilitates task- and goal-directed behavior", example traits: thinking before acting, delaying gratification, following norms and rules, planning, organizing, and prioritizing tasks), neuroticism[3] ("contrasts emotional stability and even-temperedness with negative emotionality", trait examples: feeling anxious, nervous, sad, tense), and openness[4] ("describes the breadth, depth, originality, and complexity of an individual's mental and experiential life", example traits: being imaginative, original, insightful, curious) (conceptualizations and examples taken from John et al., 2008, p. 120; for further details, see McCrae & Costa, 1997, 1999).

[2] The "Big Five" are a conceptualization which is independent from language and culture; moreover, its structure emerges across observers (self-reports, peer-reports) and methodologies (questionnaires, lexical inventories) (e.g., John & Srivastava, 1999; McCrae, 2004; Mooradian et al., 2006).

[3] Emotional instability is used as a synonym in the scientific literature.

[4] Intellect is used as a synonym in the scientific literature.

In addition to the Big Five, further well-known personality models exist. Based on work by Carl Jung (1923), Myers and Briggs developed a personality model referred to as *Myers-Briggs Type Indicator (MBTI)*. In essence, this model dichotomizes four dimensions: extraversion vs. introversion, sensation vs. intuition, thinking vs. feeling, judging vs. perceiving (Myers et al., 1998). Hence, this conceptualization comes along with $2^4 = 16$ unique personality types. Another framework is the *Eysenck personality model* which is based on two dimensions: extraversion (E) and neuroticism (N) (Eysenck, 1947). Based on high and low levels of each dimension, four personality types emerge: $E_H + N_H$: choleric type, $E_L + N_H$: melancholic type, $E_H + N_L$: sanguine type, $E_L + N_L$: phlegmatic type. Later, psychoticism was added as a third dimension, which can be divided into impulsivity and sensation-seeking (Eysenck & Eysenck, 1976). The *HEXACO model of personality* was developed by Ashton et al. (2004) and is based on the work of Costa Jr. and McCrae (1992) and Goldberg (1993). Hence, this model resembles the Big Five framework. It has six dimensions: honesty-humility, emotionality, extraversion, agreeableness, conscientiousness, and openness to experience. Thus, if compared to the Big Five, HEXACO has a sixth dimension, namely honesty-humility (with the sub-dimensions sincerity, fairness, greed avoidance, and modesty).[5] Finally, the *Revised NEO Personality Inventory (NEO PI-R)* is a model to conceptualize personality based on the Big Five. However, additionally this inventory includes six sub-factors for each trait. It follows that this model uses 30 factors to conceptualize personality. The most recent version of this inventory is referred to as NEO PI-3 (McCrae et al., 2005) and it includes the following factors: extraversion (warmth, gregariousness, assertiveness, activity, excitement seeking, positive emotions), agreeableness (trust, straightforwardness, altruism, compliance, modesty, tender-mindedness), conscientiousness (competence, order, dutifulness, achievement striving, self-discipline, deliberation), neuroticism (anxiety, hostility, depression, self-consciousness, impulsiveness, vulnerability), and openness to experience (fantasy, aesthetics, feelings, actions, ideas, values).

Trust

In situations of interpersonal interaction, trust is defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, p. 712). Moreover, it is an established fact that trustworthiness beliefs are significantly influenced by perceptions about the trustee's ability, benevolence, and integrity (e.g., Mayer et al., 1995; Rousseau et al., 1998). However, trust in machines, computers, and AI systems differs from interpersonal trust. In a recent paper, Riedl (2021) analyzed work by the research groups of David Gefen (e.g., Paravastu et al., 2014) and Harrison McKnight (e.g., Lankton et al., 2015), two well-known IS scholars with a particular focus on trust research, and concludes that more appropriate trusting beliefs in human-technology relationships are performance and functionality (instead of ability), helpfulness (instead of benevolence), and predictability and reliability (instead of integrity).[6] This conclusion is consistent with works in (e.g., Siau & Wang, 2018; Söllner et al., 2012; Thiebes et al., 2021) and outside (e.g., Hancock et al., 2011; Hoff & Bashir, 2015; Glikson & Woolley, 2020) the IS discipline.[7]

Disposition to trust (synonyms: trust propensity, interpersonal propensity to trust) also plays a major role in explaining behavioral intentions and actual behavior. McKnight et al. (1998) define disposition to trust as "a tendency to be willing to depend on others […] across a broad spectrum of situations and persons" (p. 474, 477). Disposition to trust is typically conceptualized as an antecedent of more situation-specific trust (e.g., Hoff & Bashir, 2015; McKnight et al., 1998). Thus, whether a user develops trusting beliefs and trusting intentions in, and shows actual trusting behavior toward, an IT artifact (e.g., AI system) is influenced by trust disposition. McKnight et al. (1998) further decompose disposition to trust into faith in humanity (defined as "one believes that others are typically well-meaning and reliable", p. 477) and trusting stance (defined as "one believes that, regardless of whether people are reliable or not, one will obtain better interpersonal outcomes by dealing with people as though they are well-meaning and reliable", p. 477). Complementing

[5] It is important to note that while the other five dimensions resemble the Big Five, they are not perfect matches.

[6] Based on Riedl (2021, Table 3.I.), we define as follows (in verbatim): Performance is the belief about the capability of the technology to accomplish its designated purpose. Functionality is the belief that the technology has the capability, functions, or features to fulfil the requirements. Helpfulness is the belief that the specific technology provides adequate and responsive help for users. Predictability is the belief that the technology will do what it is claimed to do without adding anything malicious on top of it. Reliability is the belief that the technology will consistently operate properly.

[7] Note that trusting beliefs in the context of automation technology and autonomous systems may also refer to purpose (defined as "the degree to which the automation is being used within the realm of the designer's intent") and process (defined as "the degree to which the automation's algorithms are appropriate for the situation and able to achieve the operator's goals" (Lee & See, 2004, p. 59). For further details, please see a recent paper by Thiebes et al. (2021) who mapped trusting beliefs (performance, etc.) and AI frameworks (e.g., OECD principles on AI) to five trustworthy AI principles (beneficence, non-maleficence, autonomy, justice, and explicability).

McKnight et al.'s (1998) work, Gefen (2000) argues that trust disposition "is not based upon experience with or knowledge of a specific trusted party […] but is the result of an ongoing lifelong experience […] and socialization […] disposition to trust is most effective in the initiation phases of a relationship when the parties are still mostly unfamiliar with each other […] and before extensive ongoing relationships provide a necessary background for the formation of other trust-building beliefs, such as integrity, benevolence, and ability" (p. 728). Consistent with these views, Mayer et al. (1995) argued that "[p]ropensity to trust is proposed to be a stable within-party factor that will affect the likelihood the party will trust […] People with different developmental experiences, personality types, and cultural backgrounds vary in their propensity to trust" (p. 715).

Framework of personality and trust in AI systems

A fundamental question is how personality is related to trust, and we focus on this important question in the context of AI systems. Moreover, it is essential how both personality and trust affect acceptance and adoption.

Propensity to trust is considered to be a specific trait of agreeableness (e.g., John & Srivastava, 1999; John et al., 2008; McCrae & Costa, 1997, 1999).[8] What follows is that the more specific personality trait propensity to trust can be deduced from the more abstract trait agreeableness. This fact is consistent with discourse in the literature which indicates that personality traits can be conceptualized on different abstraction levels, both in psychology (e.g., McCrae et al., 2005) and IS (e.g., Maier, 2012). Regarding abstraction levels of personality, John and Srivastava (1999) wrote with reference to the Big Five that "these five dimensions represent personality at the broadest level of abstraction, and each dimension summarizes a large number of distinct, more specific personality characteristics" (p. 105). Models like those from Myers-Briggs, Eysenck, or HEXACO conceptualize personality on a similarly high abstraction level. However, this advantage of broad categories also has a disadvantage, the "low fidelity" (John & Srivastava, 1999, p. 124). Decades ago it has already been argued in the personality literature that choice of abstraction levels depends on the descriptive and predictive tasks to be addressed (Hampson et al., 1987), and "[i]n principle, the number of specific distinctions one can make in the description of an individual is infinite, limited only by one's objectives" (John & Srivastava, 1999, p. 124).

[8] Mooradian et al. (2006, p. 527) argue that in personality frameworks such as Costa and McCrae's NEO framework, authors often use the term "trust", despite the fact that what is actually meant is "propensity to trust".

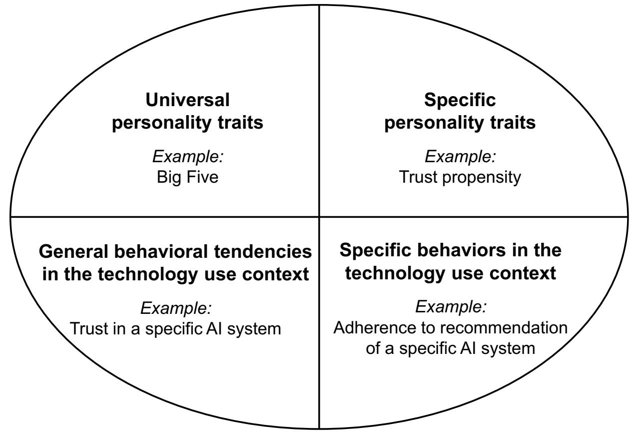


Fig. 1 Framework of user personality and trust in AI systems

Against this background, we devise a framework of personality and trust in AI systems with different abstraction levels. In personality research, it is well-established to include—in addition to the broad category on the highest abstraction level—at least one further level with more specific personality traits, which, in turn, manifest in general behavioral tendencies and more specific behaviors (e.g., Funder, 2001; John & Srivastava, 1999; McCrae et al., 2005).

Figure 1 shows a framework of personality and trust in AI systems. This framework draws upon the logic of existing conceptualizations and models at the nexus of personality and behavior (Funder, 2001; John & Srivastava, 1999; McCrae et al., 2005; Mooradian et al., 2006). In essence, the framework specifies universal personality traits (e.g., Big Five), specific personality traits (e.g., interpersonal propensity to trust), general behavioral tendencies (e.g., trust in a specific AI system such as autonomous vehicles or speech assistants like Alexa or Siri), and specific behaviors (e.g., adherence to the recommendation of an AI system in a decision-making context).

Altogether, our framework draws upon the empirically-grounded facts that (i) personality is related to behavioral tendencies and specific behaviors and (ii) universal personality traits are related to specific personality traits (e.g., Zimbardo et al., 2021). We will refer to the framework in Fig. 1 in the following sections. First, we use the distinction between universal and specific personality traits to code the scientific literature. Second, our discussion of possible future research domains is also based on this framework.

Methodology of the literature review

Literature search

In order to identify publications at the nexus of user personality and trust in the context of AI systems, we conducted a

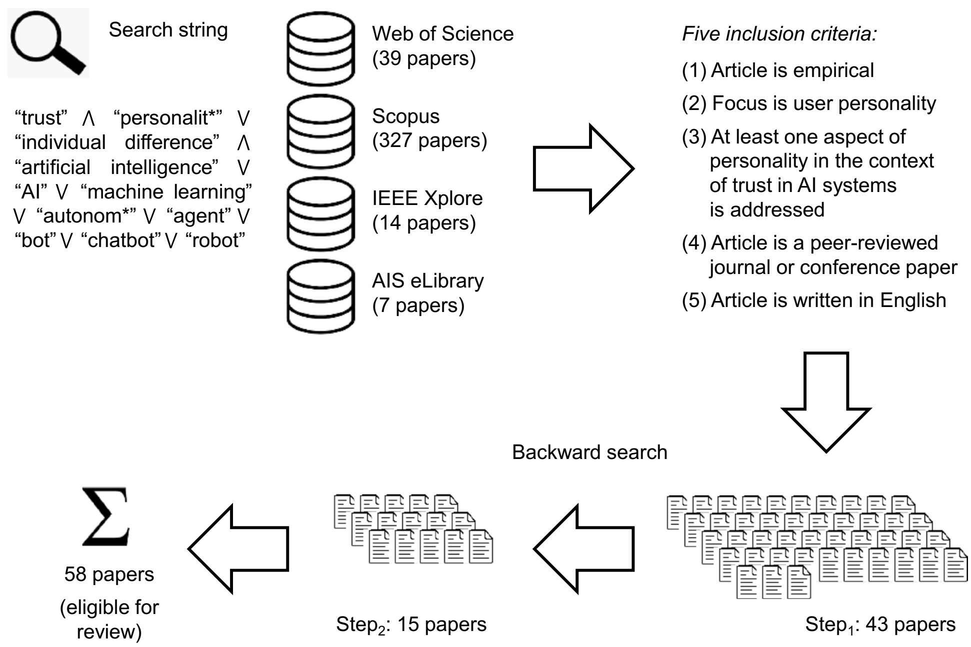


Fig. 2 Overview of the literature search process

literature search. The search process was based on existing recommendations, in particular vom Brocke et al. (2009). The search was conducted via *Web of Science, Scopus, IEEE Xplore*, and *AIS eLibrary* starting on 11/28/2021 and a last query was made on 12/31/2021. No publication year restriction was used for all searches.

Step$_1$: Because AI systems may have different manifestations, we used several keywords. Specifically, we used the following keywords: "trust", "personalit*", "individual difference", "artificial intelligence", "AI", "machine learning", "autonom*", "agent", "bot", "chatbot", and "robot" (*Web of Science, Scopus,, IEEE Xplore*). For the search in the *AIS eLibrary*, we used "trust", "personalit*", and "individual difference".[9]

This search method resulted in the following number of hits: *Web of Science* = 39 papers, *Scopus* = 327,[10] *IEEE Xplore*: 14, *AIS eLibrary*: 7. Next, we removed duplicates (as the four databases are not mutually exclusive) and read all abstracts of the remaining articles, if necessary also the full text, to identify papers eligible for our review. We applied the following inclusion criteria: (1) the article is empirical (i.e., it presents collection, analysis, and interpretation of data), (2) the focus is on user personality, and not on the personality of the AI system,[11] (3) the article deals with at least one aspect of human personality (e.g., extraversion) in the context of trust in AI systems, (4) the paper constitutes a peer-reviewed journal or conference publication, and (5) the article is written in English. After applying these inclusion criteria, 43 papers remained. Next, we read the full texts of all 43 papers and finally decided to consider them all in our review.

Step$_2$: Based on the 43 papers, we also conducted a backward search. The same five inclusion criteria were applied as in Step$_1$. Based on this procedure, we identified an additional 15 papers, all of which were ultimately considered eligible for our review.[12] Hence, the total number of papers included in our review is N = 58.[13] Fig. 2 graphically summarizes the literature search process. Appendix 1 lists the 58 references.

[9] Note that specifications such as title, topic, abstract, and keywords were not used in exactly the same way across the four databases as the search functions are not identical. Moreover, because the *AIS eLibrary* covers much fewer publication outlets than the other three databases, we did not limit our search to the context of AI in the beginning in order to organize the search more comprehensively.

[10] Note that among the high number of 327 papers identified via *Scopus*, many papers were *not* eligible for consideration in our review as this database also covers conference abstracts (which often describe research-in-progress and only have around 150 words). As a sample abstract, see Panganiban et al. (2020).

[11] For example, Hess et al. (2009) investigated how recommendation agent personality (i.e., extraversion) serves as a social technology cue to affect the perceived social presence of the agent. Thus, the focus of this study is not on user personality, but on the personality of the AI system. Hence, this paper just as many other papers with a similar focus on AI system personality were *not* considered in the present review.

[12] Note that we identified one high-quality paper published on *arXiv.org*, namely Sarkar et al. (2017). This platform is a well-known free distribution service and an open-access archive. We decided to consider this paper in our review.

[13] A forward search did not yield further relevant papers. This is plausible because half of the identified papers (i.e., 29 out of 58) were published in 2020 or 2021, and the searches itself already covered four databases, of which three are among the worldwide largest ones in the domain of the current paper (i.e., *Web of Science*, *Scopus*, *IEEE Xplore*).

Literature coding

To systematically document the existing knowledge on the relationship between user personality and trust in the context of AI systems, as well as meta-information (e.g., publication year), we extracted the following data:

(1) *General information about the paper*: (a) authors, (b) title, (c) publication year, (d) type of outlet (journal, conference), and (e) scientific discipline, in which a paper appeared. Due to the interdisciplinary nature of the current topic, we defined the following discipline categories ex-ante: (i) Computer Science, Informatics, Robotics, (ii) Ergonomics, Human Factors, (iii) Human-Computer Interaction, (iv) IS, (v) Psychology, and (vi) other.
(2) *Investigated system*: Because AI systems manifest in various forms, we used different categories, namely: (a) robot (including robot simulations), (b) autonomous vehicle (including in-vehicle agents), (c) user interface (e.g., chatbot), (d) automated agents in a security context (e.g., airport), (e) AI-driven healthcare systems, (f) autonomous system (general), and (g) other (e.g., voice assistant, intelligent tutoring). The categories were developed inductively based on the coded material.
(3) *Research method*: We grouped the empirical examinations into one of the following categories: (a) lab experiment, (b) survey study (typically conducted online), (c) online experiment, (d) lab study (in contrast to the lab experiment, this type of study does not manipulate an independent variable to systematically study the resulting effects on a dependent variable, but simply observes use of a technology in a laboratory environment). The categories were developed inductively based on the coded material.
(4) *Sample characteristics*: We documented (a) sample size, (b) mean age of the sample, (c) a sample's gender distribution, and (d) country of data collection.
(5) *Personality traits*: We documented all investigated personality traits. We distinguished between (a) universal and (b) specific traits. In the category (a), we included (i) papers which investigated all Big Five traits together and (ii) one specific Big Five trait. Moreover, we captured other broad personality conceptualizations, namely (iii) Myers-Briggs type indicator, (iv) Eysenck's model, (v) HEXACO, and (vi) NEO-PI-3. We defined these universal categories ex-ante. In the category (b), we documented all specific traits (e.g., trust propensity) which we found in the analyzed literature.
(6) *Theoretical framework*: We documented whether a paper included a theoretical framework in the form of a graphical representation (i.e., constructs and relationships). While the minimum requirement for a theory to exist is at least one independent and one dependent variable, our data analysis revealed that most papers in which a framework was available used more sophisticated theorizing. Specifically, we documented whether the investigated personality construct(s), as well as the trust construct, were conceptualized as independent, mediator, moderator, or dependent variable. Moreover, we documented all dependent variable(s) of a framework (e.g., adoption intention of an AI system).
(7) *Relationship of personality traits with trust in AI system*: We documented statistically significant relationships between universal traits, as well as specific traits, and trust in AI system. We emphasize that these relationships are typically those which the authors of a paper explicitly highlighted in the abstract and/or the discussion section of a paper. Based on this procedure, we guarantee a focus on the most important results from the perspective of the papers' authors.

Results

We structure the presentation of our results into two subsections. We start with the descriptive results (i.e., point (1) to (4) as listed in the previous literature coding section), followed by our findings on the investigated personality traits and underlying theoretical models, point (5) to (7) in the list.

Descriptive results

Appendix 1 lists the 58 references. Appendix 2 summarizes the major characteristics of these 58 papers and their underlying studies. The 58 papers report 64 studies. Based on the information in Appendix 1 and 2, Table 1 summarizes descriptive statistics of our review. Specifically, we report publication year, the type of outlet in which the papers are published, and the scientific discipline in which a paper appeared.[14] Moreover, we document the investigated technology, research method, and country of data collection. For further sample characteristics (i.e., sample size, mean age of the sample, and rate of females in the sample) we refer the reader to Appendix 2.

[14] Coding of the 58 papers regarding scientific discipline was a straightforward process. However, with respect to the Hawaii International Conference on System Sciences (HICSS) we decided to assign this publication venue to the IS discipline (and not to Computer Science, Informatics, and Robotics). The reason for this decision is that that this conference focuses on "Information Technology *Management*" (see https://aisel.aisnet.org/hicss/, italics added).

Table 1 Descriptive statistics

Metric	Findings	
Publication year (N = 58 papers)	08	2
	09	0
	10	0
	11	2
	12	2
	13	2
	14	0
	15	5
	16	2
	17	4
	18	4
	19	6
	20	17
	21	12
Publication outlet (N = 58 papers)	Conference proceedings	30
	Journal	27
	arXiv	1
Scientific Discipline (N = 58 papers)	Computer Science, Informatics, Robotics	18
	Ergonomics, Human Factors	13
	Human-Computer Interaction	11
	Information Systems (IS)	8
	Psychology	4
	Other (e.g., Communication, Sociology)	4
Investigated system (N = 59 studies for which the system is reported)	Robot (including robot simulations)	20
	Autonomous vehicle (including in-vehicle agents)	12
	User interface (e.g., chatbot)	11
	Automated agent in security context (e.g., airport)	7
	AI-driven healthcare systems	3
	Autonomous system and AI (general)	3
	Other (e.g., voice assistant, intelligent tutoring)	3
Research method (N = 60 studies for which the method is reported)	Laboratory experiment	22
	Survey study	18
	Online experiment	14
	Laboratory study	6
Country (data collection) (N = 60 studies for which the country is reported)	USA	30
	Mix of different countries	8
	Germany	6
	United Kingdom	4
	Canada	3
	Italy	2
	Australia	2
	Japan	2
	China	1
	Korea	1
	South Africa	1

Findings on personality traits

Table 2 summarizes our findings on the investigated personality traits, separated into universal and specific traits.

Regarding the universal traits, we observed that 18 out of the 58 papers studied the Big Five. Moreover, in 11 further papers, single Big Five traits were studied. While extraversion was studied in 5 papers, neuroticism was studied in 3, agreeableness in 2, and openness in 1 paper, respectively. No article had a focus on conscientiousness alone. Moreover, the results show that other general personality models only play a minor role in the extant literature. Specifically, the Myers-Briggs and Eysenck models, as well as HEXACO and NEO-PI-3, have only been applied once each.

Table 2 further indicates that we identified 33 specific personality traits.[15] We list all traits which were studied in at least two papers (i.e., 12 traits).[16] We found that trust propensity (i.e., the disposition to trust other people) and dispositional trust in automation/machines are the most studied

[15] Note that we harmonized terminology across the 58 papers in our analyses and tables. For example, we use the term “trust propensity” despite the fact that synonyms like “interpersonal propensity to trust”, “disposition to trust”, “dispositional trust”, “dispositional interpersonal trust”, “dispositional trust in humans”, or “trust disposition” are used in the literature.

[16] In Appendix 2 (see column “Specific traits”) the reader can find the 21 traits which were only examined in one of the 58 papers.

Table 2 Statistics on the personality traits

Traits	Findings	
Universal personality traits (N = 58 papers)	Big Five (together)	18
	Extraversion	5
	Neuroticism	3
	Agreeableness	2
	Openness	1
	Conscientiousness	0
	Myers-Briggs	1
	Eysenck	1
	HEXACO	1
	NEO-PI-3	1
Specific personality traits (N = 58 papers)	Trust propensity	9
	Dispositional trust in automation/machines	9
	Locus of control	7
	Attitude towards computers/robots	5
	Self-efficacy	5
	Technological innovativeness	4
	Need for cognition	3
	Perfect automation expectation	3
	Propensity to take risks	3
	Sensation seeking	3
	Need for interaction	2
	Self-esteem	2
	Other (e.g., computer anxiety, thrill seeking)	21

specific personality traits (in 9 papers each). Also, locus of control plays a significant role in the literature (studied in 7 papers). Locus of control is the degree to which people believe that they have control over the outcome of events (as opposed to external forces beyond their influence); this locus can either be internal (a person believes that he or she can control their own life) or external (the belief that factors which cannot be influenced control life) (Rotter, 1966, 1990). Both attitude towards computers/robots and self-efficacy (the degree to which people believe that they can accomplish a particular task or activity, Bandura, 1977, Gist, 1987) were studied in 5 papers each. The following traits were examined in three papers each: need for cognition ("an individual's tendency to engage in and enjoy effortful cognitive endeavors", Cacioppo et al., 1984, p. 306), perfect automation expectation ("expectations that the automated aid [e.g., autonomous vehicle] will perform with near-perfect reliability", Merritt et al., 2015, p. 740),[17] propensity to take risks, and sensation seeking ("the tendency to seek novel, varied, complex, and intense sensations and experiences and the willingness to take risks for the sake of such experiences", Choi & Ji, 2015, p. 694). Finally, need for interaction and self-esteem were studied in 2 papers each. Need for interaction is a person's preference to stay in touch with other people during the use of a service or application (Dabholkar, 1992). Self-esteem is defined as "the level of global regard one has for the self as a person" (Harter, 1993, p. 88).

Regarding the availability of theoretical frameworks, we found graphical representations with constructs and relationships in 19 out of the 58 papers.[18] Table 3 summarizes the results of our analyses. As shown in Table 3, we found that 18 out of the 19 papers conceptualize personality traits as independent variable. Another major result is that trust is conceptualized as independent variable in 3 papers only. However, trust is much more often conceptualized either as mediator (7 times) or as dependent variable (11 times). Table 3 further shows that personality traits are rarely conceptualized as mediator (2 times) or moderator (3 times). Also, we observe that in studies in which trust is not the dependent variable, often behavioral intention to use an AI system is used as outcome variable.

Regarding the relationship between personality traits and trust in AI systems, a first notable result was that many of the hypothesized relationships in the analyzed papers were statistically *not* significant. This concerns both the

[17] Note that this expectation is related to the concept of perfect automation schema (PAS), which has two dimensions: first, high expectations for automation performance, and second, the fact that users hardly forgive computers and machines when they make mistakes. For further details, see Dzindolet et al. (2002).

[18] We only focused on frameworks with direct relevance for the present article (i.e., the framework has at least one personality construct and one trust construct).

Table 3 Analysis of the theoretical conceptualization of personality and trust (N = 19 papers in which graphical representations of theoretical frameworks are provided)

Reference	Pers = IV	Trust = IV	Pers = MED	Trust = MED	Pers = MOD	Trust = MOD	Pers = DV	Trust = DV	DV
Bawack et al. (2021)	x			x					Customer experience performance
Choi and Ji (2015)	x	x							Behavioral intention
Cohen & Sergay (2011)	x	x							Behavioral intention
Elson et al. (2018)				x	x				Decision
Elson et al. (2020)	x							x	
Handrich (2021)	x			x					Adoption intention
Hanna & Richards (2015)	x							x	
Hegner et al. (2019)	x	x		x					Adoption intention
Kim et al. (2020)	x							x	
Kraus et al. (2020a)	x		x					x	
Kraus et al. (2020b)	x							x	
Matthews et al. (2020)	x							x	
Merritt et al. (2013)	x				x			x	
Miller et al. (2021)	x			x				x	
Schaefer & Scribner (2015)	x			x					Performance
Sharan & Romano (2020)	x							x	
Youn & Jin (2021)	x		x		x			x	Intention to visit, satisfaction
Zhang et al. (2020)	x			x					Behavioral intention
Zhou et al. (2020)	x							x	
Sum	18	3	2	7	3	0	0	11	

IV independent variable, *MED* mediator variable, *MOD* moderator variable, *DV* dependent variable, *Pers* Personality

Big Five traits and the more specific traits. However, we also observed relationships which turned out to be stable across the analyzed papers.

With respect to the *Big Five*, we found the following results:

- Agreeableness: 9 papers found a positive relationship with trust in AI systems (Bawack et al. 2021, Böckle et al. 2021, Chien et al. 2016, Ferronato & Bashir 2020a, Huang et al. 2020, Kraus et al. 2020a, Lyons et al. 2020, Müller et al. 2019, Rossi et al. 2018).
- Openness: 8 papers found a positive relationship with trust in AI systems (Aliasghari et al. 2021, Antes et al. 2021, Böckle et al. 2021, Elson et al. 2020, Ferronato & Bashir 2020a, Oksanen et al. 2020, Schaefer & Straub 2016, Zhang et al. 2020).
- Extraversion: 5 papers found a positive relationship with trust in AI systems (Böckle et al. 2021, Haring et al. 2013, Kraus et al. 2020a, Merritt & Ilgen 2008,

Müller et al. 2019), while 1 paper found a negative relationship (Ferronato & Bashir 2020a). One further study found that the relationship between extraversion and trust is more complex. Elson et al. (2018) write: "Individuals high in extraversion initially rated their trust in the agent as higher than those low in extraversion. During [... further interaction] rounds, the agent gave a recommendation contrary to what was expected in highly confident conditions [...] trust in the agent increased compared to the previous round for individual's low in extraversion while trust in the agent decreased compared to the previous round for individuals high in extraversion [...] extraverts appeared to have their trust most dramatically affected by the conditions where the agent gave an incorrect response in the highly confident condition" (p. 436). What follows is that the level of trust in an AI system is influenced by an interaction of the degree of extraversion and whether an expectation regarding decision outcome is fulfilled.
- Conscientiousness: 3 papers found a positive relationship with trust in AI systems (Bawack et al. 2021, Chien et al. 2016, Rossi et al. 2018), while 2 papers found a negative relationship (Aliasghari et al. 2021, Oksanen et al. 2020).
- Neuroticism: 3 papers found a negative relationship with trust in AI systems (Kraus et al. 2020a, Sharan & Romano 2020, Zhang et al. 2020).

With respect to *specific personality traits*, we observed a *positive relationship* of the following constructs with trust in AI systems:

- trust propensity (Aliasghari et al. 2021, Huang & Bashir 2017, Rossi et al. 2018),
- perfect automation expectation (Lyons & Guznov 2019, Lyons et al. 2020, Matthews et al., 2020),
- dispositional trust in automation/robots (Kraus et al. 2020a, Merritt & Ilgen 2008),
- self-efficacy (Handrich 2021, Oksanen et al. 2020),
- technological innovativeness (Handrich 2021),
- sensation seeking (Zhang et al. 2020), and
- self-esteem (Kraus et al. 2020a).

Moreover, we observed a *negative relationship* between trust in AI systems and the following traits:

- internal locus of control (Chiou et al. 2021, Sharan & Romano 2020),
- negative attitude towards computers/robots (Matthews et al., 2020, Miller et al. 2021), and
- propensity to take risks (Ferronato & Bashir 2020b).

To sum up, our review of the relationship between universal personality traits and trust in AI systems indicates that for three factors out of the Big Five (agreeableness, openness, extraversion) high values positively affect trust, while high neuroticism values negatively affect trust. With respect to conscientiousness, evidence is mixed and hence no clear statement on the relationship between this personality trait and trust in AI systems can be made based on the current research status. However, it is critical to consider that the reported results are only valid with the meaning of "all-other-things-being-equal". As an example, an individual with high agreeableness trusts an AI system more than an individual with low agreeableness, but only if the two individuals do not differ in the other four personality traits (we reflect on this finding in the following Discussion section). Regarding the relationship between specific personality traits and trust in AI systems we found that research more frequently established a positive relationship (seven factors) than a negative relationship (three factors). However, in contrast to Big Five research the total number of available studies on most specific personality traits is lower and hence the reported results should be replicated.

Discussion and future research directions

We structure our discussion into two sub-sections. We start with a discussion of the descriptive results (as summarized in Table 1), followed by a discussion of the findings on personality traits (as summarized in Tables 2 and 3). This discussion also includes the identification of unexplored research areas and an outline of possible directions for future research.

Descriptive results

Regarding *publication year*, we identified the first studies in 2008 (Cramer et al. 2008; Merritt & Ilgen 2008). Since then, only a limited number of studies were published up until the recent past. Notably, while 29 out of the 58 identified papers were published in the period 2008-2019, the remaining 29 studies were published in 2020 and 2021. Thus, in the very recent past we observe a significant increase in empirical research at the nexus of user personality and trust in the context of AI systems, signifying the sharply rising academic interest in the topic.

Regarding *publication outlet*, we found almost a balance between conference papers (30 papers) and journal publications (27 papers). Considering that in several of the disciplines in which the investigated studies were published, conference publications are of high scientific value (e.g., computer science, HCI), we consider this balance as a "good sign", predominantly because articles in conference proceedings guarantee that research results are quickly available (while review processes for journals typically can take much longer). Yet, this statement should not be interpreted as an argument against journal articles. Rather, the found balance is the state which is also desirable in the future.

Regarding *scientific discipline*, we identified computer science, informatics, and robotics as the dominant group (18 papers), followed by ergonomics and human factors (13 papers), and HCI (11 papers). Information systems (IS, 8 papers) and psychology (4 papers) contributed less frequently to the current body of research. A major explanation for this observation is that the dominating disciplines have already concentrated research efforts on personality, trust, and intelligent systems in the past decade, while IS and psychology did not. However, considering that recent AI articles published in mainstream IS outlets (e.g., Berente et al., 2021, *MIS Quarterly*) explicitly refer to trust as a "managerial issue" (p. 1440), and that three of the four identified psychology papers recently appeared in a widely visible publication outlet, namely in *Frontiers in Psychology* (Kraus et al. 2020b, Miller et al. 2021, Oksanen et al. 2020), we foresee that the disciplines of IS and psychology will also contribute more research in the future. Considering that it is unlikely that the currently dominating disciplines will slowdown in their publication output (Matthews et al., 2021), it is definitive that research at the nexus of user personality and trust in the context of AI systems will most likely experience an upward tendency in the future.

Regarding *investigated systems*, our analyses revealed that robot research (including robot simulations) are by far dominant (20 papers), followed by autonomous vehicle studies (12 papers) and user interface examinations (e.g., chatbots) (11 papers). Further domains identified by our review are automated agents in the security context (e.g., airports) (7 papers), AI-driven healthcare systems (3 papers), as well as autonomous systems and AI in general (3 papers), along with some other systems such as voice assistants (3 papers). At least to some extent, this finding reflects the historical fact that computer science, informatics, and robotics, as well as ergonomics and human factors, heavily focused their research on robots (e.g., Rossi et al. 2020) and autonomous vehicles (e.g., Tenhundfeld et al. 2020). However, as a consequence of the foreseen rise of IS and psychological research, AI systems embedded into user interfaces (e.g., automated decision-support in the business context), as well as intelligent health systems and home systems (e.g., speech assistants), will likely gain in importance in the future (see, for example, a recent *MIS Quarterly* special issue, No. 3, September 2021). We explicitly make a call for studies in these under-researched domains.

Furthermore, regarding investigated systems we observed that the 58 papers remained vague in the reporting of the studies' task instructions. It follows that an inherent limitation of the analyzed articles is that it is *not* clear if the identified influences of personality traits on trust in AI systems are indeed specific to AI systems or hold for IT systems in general. Considering the idiosyncrasies of AI systems and their possible influences on trust (e.g., Siau & Wang, 2018; Thiebes et al., 2021), future studies should report their task instructions in detail (ideally in verbatim) in order to be better able to disentangle the specific AI effects from the more general IT effects. Such a description of task instructions should also include the definition of AI, which was provided to the study participants, if provided at all.[19]

Regarding applied *research method*, we observe that except for some qualitative interviews which are reported as a "by-product" of larger quantitative studies (e.g., Sarkar et al., 2017),[20] the current body of research almost entirely used quantitative methods. The dominant method is the laboratory experiment (22 papers), followed by survey studies (18 papers), online experiment (14 papers), and a few laboratory studies without experimental manipulation (6 papers). We see three major implications.

First, despite the uncontested value of quantitative research which is rooted in the positivist paradigm (e.g., Chen & Hirschheim, 2004), other epistemological positions and methods, particularly interpretivism and qualitative methods (e.g., Walsham, 1995), should become more important in the future. Such an interpretivist stance is often deeply rooted in a hermeneutic tradition, thereby being of a fundamentally idiographic nature. Such research (e.g., interpreting interview data on individuals' perceptions of their interaction with AI systems and theory building), therefore, has the objective of determining "patterns" *and* "richness in reality" (Mason et al., 1997, p. 308).

Second, detailed analysis of the identified online experiments revealed that these studies typically used Amazon Mechanical Turk (MTurk) for data collection. MTurk is a platform that offers access to a geographically dispersed set of respondents. While such a sample "can be more representative and diverse than locally collected samples […] respondents tend to be relatively young, digitally savvy adults" (Antes et al. 2021, p. 3). It follows that future research should replicate existing research findings with more representative populations. This call for future studies

[19] The author would like to thank one anonymous reviewer for outlining this important point. To substantiate the author's manual check of this finding, a formal search was made in all 58 papers based on the terms "task instruction" and "task description". Two hits were found for "task instruction" (Harriott et al., 2018, Sharan & Romano 2020) and one hit was found for "task description" (Harriott et al., 2018). However, these three text passages also do not specify the task in a way so that the reader exactly knows the participants' notion of the investigated system.

[20] Sarkar et al. (2017) conducted a lab experiment and as a complement collected interview data.

is substantiated by the fact that attitude towards AI systems (e.g., Nam, 2019), trust (e.g., Sutter & Kocher, 2007), and personality (e.g., Ashton & Lee, 2016) are related to age. As indicated in Appendix 2 (see table "mean age (years)"), we identified information on participant age in 60 out of the 64 studies. The average age of participants, or the dominating age group in studies in which the average age was not reported, is >40 years in only 5 studies (Matsui 2021: 44.6 y, Rossi et al. 2020: 61 y, Sorrentiono et al. 2021: 83.33 y, Voinescu et al. 2018: 67.52 y, Youn & Jin 2021: 40.8 y). Therefore, the findings of the current review are predominantly valid for younger age groups. Future research should be based more often on older people. This call is substantiated by the fact that older people are increasingly concerned by AI systems, as signified by the example of socially assistive robots (e.g., Sorrentiono et al. 2021).

Third, during the paper screening and selection process (see Fig. 2) we encountered a paper by Mühl et al. (2020) who used neurophysiological measurement (i.e., skin conductance) as a complement to self-reports. However, because this paper studied passengers' preferences and trust while being driven by a human driver and an intelligent vehicle without any focus on user personality, we did *not* consider this article in our body of analyzed literature. It follows that no single article in our sample of N = 58 papers used neurophysiological measurement. Importantly, research has demonstrated "associations between personality traits and the structure and function of the nervous system" (Funder, 2001, p. 206). Thus, we consider the application of neuroscientific measurement techniques as a promising methodological avenue for future studies. In this respect, the study by Mühl et al. (2020) could be a promising starting point. This call for the use of neuroscience approaches is substantiated by two research developments: first, the increasingly important role of neuroscience approaches in IS research, referred to as NeuroIS (e.g., Dimoka et al., 2012; Riedl & Léger, 2016), and second, the increasing evidence that personality traits have a strong evolutionary, affective, and hence biological component (e.g., Montag et al., 2016; Montag & Panksepp, 2017).

Regarding *country of data collection*, the most striking result is that in 30 out of 60 studies in which the country is reported, data gathering took place in the United States. Then, the category "mix of different countries" follows. This category comprises MTurk studies in which subjects from more than one country participated. Six studies collected data in Germany, followed by the UK (4 studies), Canada (3 studies), Italy, Australia, and Japan (2 studies each), as well as three further countries in which one study was conducted (China, Korea, South Africa). The major implication of this finding is that more future studies should be carried out outside the United States. This call is substantiated by evidence showing that personality and culture are related (e.g., Leung & Cohen, 2011; McCrae, 2000). One study in our sample investigated the relationship between trust attitudes towards automated digital technologies, Hofstede's cultural dimensions, and the Big Five personality traits (Chien et al. 2016). Reflecting on their results, Chien et al. write: "the U.S. population had the highest trust score and Turkish group scored the lowest, with Taiwanese population falling in between [...] Evaluations of the inter-relational aspects of personality and general trust showed that an individual with a high trait of agreeableness or conscientiousness had increased trust in automation" (p. 845). Thus, culture should be a construct of interest in future research. This call for future studies is substantiated by long existing calls for more cultural studies in a seminal psychology paper. Funder (2001) writes: "[the] direction is to try to distinguish between the psychological elements that are shared by all cultures (etics) and those that are distinctive to particular cultures (emics) […] The big five have been offered as possible etics" (p. 203). Against this background, fruitful avenues for future research could be to examine the possible relationship between culture and personality traits in the context of trust in AI systems.

A seminal paper by Hofstede and McCrae (2004) could serve as a starting point.[21] They found that several culture dimensions are related to the Big Five; specifically, they report the following statistically significant correlations[22]: *power distance* ("the extent to which the less powerful members of organizations and institutions […] accept and expect that power is distributed unequally") correlates with extraversion (−), conscientiousness (+), and openness (−); *uncertainty avoidance* ("a society's tolerance for ambiguity [… indicating] to what extent a culture programs its members to feel either uncomfortable or comfortable in unstructured situations [… which] are novel, unknown, surprising, and different than usual") correlates with neuroticism (+) and agreeableness (−); *individualism* ("the degree to which individuals are integrated into groups [… and in] individualist societies, the ties between individuals are loose: everyone is expected to look after himself or herself and his or her immediate family") correlates with extraversion (+); finally, *masculinity* ("the distribution of emotional roles between the sexes […] women's values differ less among societies than men's values [… and] men's values vary along a dimension from very assertive and competitive and maximally different from women's values on one side to modest and caring and similar to women's values on the other") correlates with

[21] The platform *www.hofstede-insights.com* defines culture as "the collective mental programming of the human mind which distinguishes one group of people from another" and it offers a feature to compare countries based on various culture dimensions.

[22] Definitions of culture dimensions are taken in verbatim from Hofstede and McCrae (2004, pp. 62-63). The signs "+" and "-" indicate positive or negative correlations.

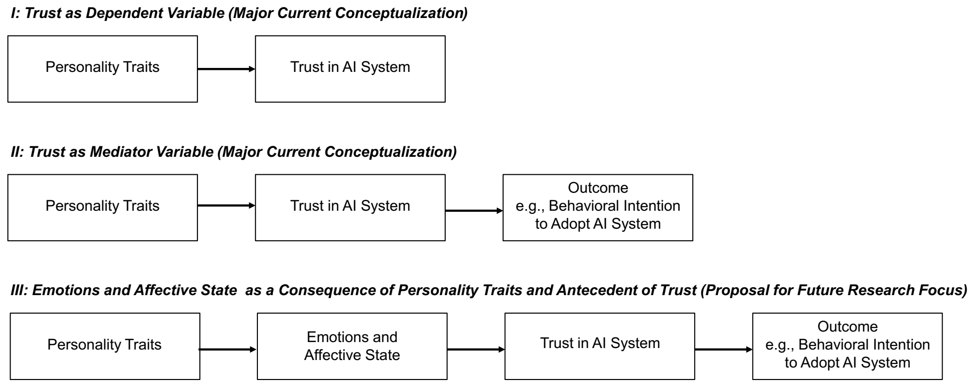


Fig. 3 Two dominating theoretical conceptualizations of personality and trust in the context of AI system adoption and proposal for future research focus

openness (+), neuroticism (+), and agreeableness (−). This existing knowledge should be used in future studies on trust in AI systems.

Findings on personality traits

Regarding the *universal personality traits*, we found that the Big Five were studied extensively, while other general personality models were not. Thus, future research should consider these other models more frequently. In particular, we recommend future studies based on HEXACO (Ashton et al., 2004) and NEO-PI-3 (McCrae et al., 2005), as these models are both related to the Big Five and relatively novel frameworks in the personality literature (if compared to Myers-Briggs' and Eysenck's models). In our sample of analyzed papers, Müller et al. (2019) used HEXACO to study personality and trust in the context of human-chatbot interaction. Moreover, Rossi et al. (2020) used NEO-PI-3 to examine human interaction with socially assistive robots in a simulated home environment. These two studies may serve as a starting point for future research.

Regarding the *theoretical conceptualization* of personality and trust in studies on adoption and interaction with AI systems (see Table 3), our review reveals two dominating causal chains. First, personality is conceptualized as an independent variable and trust in the AI system as a dependent variable (e.g., Elson et al. 2020). Second, personality is conceptualized as an independent variable, trust in the AI system as a mediator variable, and another outcome such as behavioral intention to adopt the AI system is conceptualized as a dependent variable (Handrich 2021). Figure 3 graphically summarizes these two theoretical mechanisms, and we added a third one. This additional mechanism describes a fruitful future research focus in which emotions and affective state is conceptualized as a consequence of personality traits and as an antecedent of trust in AI system, which in turn influences behavioral intention. A major motivation for consideration of this additional mechanism is the fact that emotions and affect have become critical constructs in IS research in general (e.g., Zhang, 2013) and in AI system adoption research in particular (e.g., Kraus et al. 2020b, Miller et al. 2021).

The Merriam-Webster Dictionary defines *affect* as "a set of observable manifestations of an experienced emotion: the facial expressions, gestures, postures, vocal intonations, etc., that typically accompany an emotion".[23] Thus, emotions are closely related to physiological processes, followed by observable consequences of these processes—the affect. Interestingly, despite a few notable exceptions such as Kraus et al.'s (2020b) work on automated driving or Miller et al.'s (2021) work on human-robot interaction, emotions and affect hardly play a role in the analyzed literature. Thus, the role of emotions and affect has hardly been established in the present research context.

In Fig. 3 (bottom), we suggest positioning emotions and affective state as a consequence of personality traits and as an antecedent of trust, which in turn influences further downstream variables, such as behavioral intention to adopt an AI system. This positioning is consistent with the theoretical frameworks in Kraus et al. (2020b) and Miller et al. (2021). For example, Miller et al. (2021) write: "Besides user dispositions, users' emotional states during the familiarization with a robot are a potential source of variance for robot trust. As the experience of emotional states has been shown to be considerably affected by personal dispositions, this research proposes a general mediation mechanism from the effects of user dispositions on trust in automation by user states […] focus[ing] on state anxiety as a specific affective state, which is expected to explain interindividual differences in trust in robots […] State anxiety is defined as "subjective, consciously perceived feelings of apprehension and tension, accompanied by or associated with activation or arousal of the autonomic nervous system" […]" (p. 5). Future empirical research should directly test this causal

[23] https://www.merriam-webster.com/dictionary/affect (accessed on January 16, 2022).

chain.[24] Methodologically, we recommend two things: first, to complement existing personality measurement instruments (e.g., Big Five) with the Affective Neuroscience Personality Scales (ANPS) (Montag et al., 2021; Orri et al., 2017; Reuter et al., 2017), and second, to consider measuring emotions based on neurophysiological measurement (NeuroIS). The major argument for this suggestion is that trusting beliefs in and attitudes towards AI systems are not only influenced by conscious perceptions and thoughts (which would imply the use of self-report measurement). Rather, they are also influenced by unconscious perceptions and information processing which cannot be captured based on self-reports alone (e.g., Dimoka et al., 2012; Riedl & Léger, 2016; vom Brocke et al., 2020). Thus, future research should consider measuring emotions, and measurement should include both self-report and neuroscience instruments.

As outlined in detail in the Results section, we examined the *Big Five's influence on trust in AI systems*. In essence, overwhelming evidence shows that both agreeableness and openness positively affect trust. Moreover, while the evidence is less compelling if compared to agreeableness and openness, there is still a clear tendency that extraversion also positively affects trust. Regarding neuroticism, the findings are also clear-cut. Less neurotic people exhibit more trust in AI systems. However, this finding is based on fewer studies if compared to the other Big Five traits.

Finally, regarding the influence of conscientiousness on trust in AI systems the evidence is mixed. While some studies (Bawack et al. 2021, Chien et al. 2016, Rossi et al. 2018) found a positive relationship, other studies found a negative relationship (Aliasghari et al. 2021, Oksanen et al. 2020). For example, based on the context of a domestic robot for food preparation, Aliasghari et al. (2021) found that more conscientious people trusted the robot less, and hence were more likely to rely on themselves as cook or a restaurant. In contrast, another study which was carried out in the context of AI-based voice shopping (e.g., Amazon Echo) found a positive relationship between conscientiousness and trust (Bawack et al. 2021). What follows is that variance in AI system (domestic robot vs. AI-based voice assistant) and/or in the task (food preparation vs. shopping), may have caused the different relationships between the personality trait conscientiousness and trust. Thus, it is a fruitful scientific endeavor to study more systematically the influence of specific AI systems and/or the technology-supported tasks (as moderators) on the relationship between personality (independent variable) and trust (dependent variable).

We highlight that trust can vary strongly depending on what kind of AI system a user is dealing with. For example, highly critical AI systems such as medical AI systems would be harder to trust than non-critical systems such as voice assistants or recommender systems (due to the potential harm they might cause in case of a malfunction or a poor system design). Thus, despite the fact that the current research status suggests a direct relationship between the personality traits agreeableness (positive), openness (positive), extraversion (positive), and neuroticism (negative) and trust in AI systems, more research is necessary to determine whether these findings hold in the context of highly critical AI systems. Importantly, no paper in our sample of 58 empirical articles chose the context of highly critical AI systems like those in the medical domain. Thus, today we do not know whether the current results may be generalized to this context. Also, other boundary conditions should be considered in future research, including user experience and computer self-efficacy (Matthews et al., 2021).

Moreover, the kind of collaboration process should be considered in future studies. In essence, decision automation should be distinguished from scenarios where AI acts as a decision support system. Our analysis of the extant literature (see Appendix 2, column "Context") indicates that the available literature predominantly dealt with decision support systems rather than complete decision automation. Thus, the influence of personality traits on trust in AI systems which fully automate decisions should be studied more frequently in the future. Several researchers are aware of this necessary distinction; however, they have not yet considered this distinction in their research designs. As an example, Tenhundfeld et al. (2020) studied trust in automated parking based on a Tesla vehicle. However, they highlight that their research focus was "*partially* automated parking" (p. 194, italics added) and not complete decision automation.[25]

Despite the fact that current knowledge on the relationship between the Big Five traits and trust in an AI system is substantial, the practical implication of this knowledge is limited. Every human can be characterized as a configuration of the Big Five values. In the simplest case in which each trait is assumed to be either high or low we have $2^5 = 32$ personality types. However, the existing body of literature does *not* focus on investigation of these types' influence on trust. Rather, the extant literature established a trait's individual relationship with trust in a specific AI system. This kind of knowledge is limited as it only allows for trust prediction for some configurations. Based on the present review's knowledge, a person with profile$_1$ would

[24] Kraus et al. (2020b) theoretically suggest *and* empirically confirm the following causal chain: personality traits as independent variable, emotional state as mediator, and trust as dependent variable. Replication studies should be conducted to better establish this mechanism.

[25] The author would like to thank one anonymous reviewer for outlining this important point.

trust an AI system more than a person with profile_2 (initial letters of the Big Five, "↑" indicates a high level, while "↓" indicates a low level): $\text{profile}_1 = A\uparrow + O\uparrow + E\uparrow + N\downarrow + C\uparrow$, $\text{profile}_2 = A\downarrow + O\downarrow + E\downarrow + N\uparrow + C\uparrow$. However, what would be the prediction for the following configuration: $\text{profile}_3 = A\uparrow + O\downarrow + E\uparrow + N\uparrow + C\uparrow$? Based on the insights presented in the extant literature, it would be hardly possible to make a sound prediction. However, one paper in the analyzed literature used latent profile analysis to study trust in chatbots (e.g., Alexa). Based on this technique and the HEXACO model, Müller et al. (2019) not only identified extraversion, agreeableness, and honesty-humility as the three most relevant traits for trust prediction, but also determined *specific personality profiles with predictive power for trust*: "introverted, careless, distrusting user", "conscientious, curious, trusting user", and "careless, dishonest, trusting user" (p. 35). More research of this sort is necessary in the future.

Regarding *specific personality traits*, we found some clear-cut results. First, trust propensity (Aliasghari et al. 2021, Huang & Bashir 2017, Rossi et al. 2018), perfect automation expectation (Lyons & Guznov 2019, Lyons et al. 2020, Matthews et al., 2020), and dispositional trust in automation/robots (Kraus et al. 2020a, Merritt & Ilgen 2008) positively affect situational trust in a specific AI system. Moreover, we found the following traits to also positively affect this trust: self-efficacy, technological innovativeness, sensation seeking, and self-esteem. However, as indicated in the Results section, because only a very limited number of studies have examined these constructs, future research should replicate these findings before more definitive conclusions can be drawn.

Two specific traits turned out to be negative predictors of trust in AI systems (at least two independent studies observed this effect for both factors): negative attitude towards computers/robots (Matthews et al., 2020, Miller et al. 2021) and internal locus of control (Chiou et al. 2021, Sharan & Romano 2020). It is highly plausible that people with a negative attitude towards computers and robots generally do exhibit little situational trust in a specific AI system. Regarding locus of control, evidence shows that people who believe that they can control their own lives (rather than being controlled by external forces) exhibit less trust towards AI systems. This finding is also plausible because a strong belief in one's own ability to control life may come along with less trust in everything else which could control one's life, either other humans or technological artifacts like AI-based systems. The study by Choi & Lee (2015) provides another interesting rationale for this finding; based on an autonomous vehicle context they write: "locus of control […] selected as driving-related personality factor […] External locus of control significantly influenced behavior. This result demonstrated that someone who experiences difficulty in driving, such as older adults, has greater intention to use the autonomous vehicles" (p. 699). It follows that people who believe that they do not have sufficient capabilities and skills to interact with or handle an intelligent system (= no internal locus of control) do trust an intelligent system *because* there is no alternative.

Overall, a major conclusion of the current review is that today only a limited number of specific personality traits were studied sufficiently so that solid knowledge exists (i.e., trust propensity, perfect automation expectation, and dispositional trust in automation/robots positively affect trust in AI systems, while negative attitude towards computers/robots and internal locus of control have a negative effect). However, more research should be conducted on the remaining traits which we found in our literature analysis (as described in detail in the column "Specific traits" in Appendix 2) as well as further traits which are described elsewhere (Matthews et al., 2021).

Our call for more research on specific personality traits is substantiated by the fact that the Big Five do *not* subsume "all there is to say about personality" (Funder, 2001, p. 200). As an example, consider an authoritarian personality (Adorno et al., 1950). Funder (2001) indicates that such a personality "would be high on conscientiousness and low on agreeableness and openness […] but much would be lost if we tied to reduce our understanding of authoritarianism to these three dimensions" (p. 201). It follows that while specific personality traits are definitely related to more abstract traits such as the Big Five, they feature additional information. Therefore, we reiterate for the domain of trust in AI systems what Funder (2001) already declared more than two decades ago with reference to the Big Five, namely that researchers should not be "seduced by convenience [… and] act as if they can obtain a complete portrait of personality by grabbing five quick ratings" (p. 201).

As indicated in Table 3 and Fig. 3 (see "Trust as Mediator"), some studies used behavioral intention, or adoption intention, as dependent variable. Thus, these studies provide first insights on how personality analysis can be used to improve the adoption process of AI systems. One study, for example, indicates that system transparency (e.g., "I believe that I can form a mental model and predict future behavior of the autonomous vehicle."), technical competence (e.g., "I believe that I can depend and rely on the autonomous vehicle."), and situation management (e.g., "I believe that the autonomous vehicle will provide adequate, effective, and responsive help.") positively affected trust, which, in turn, increased behavioral intention to use an autonomous vehicle (Choi & Ji, 2015). In another study, it is reported that need for interaction, self-efficacy, technological innovativeness, and novelty seeking positively affected trust, which, in turn, reduced innovation resistance and increased adoption intention of intelligent personal assistants like Alexa (Handrich

2021). These and further studies (Cohen & Sergay 2011, Hegner et al. 2019, Zhang et al. 2020) constitute a starting point for future research that has the goal to develop recommendations on how to improve the adoption process of AI systems based on user personality knowledge. As signified by the examples above, current research suggests—if considered as collective evidence—that the *interaction* of personality traits (either the Big Five or specific traits like those investigated by Handrich 2021) with system properties (like those investigated by Choi & Ji, 2015) determines trust. In addition to system properties, task demands (e.g., workload in partly automated tasks) should also be considered. This need is substantiated by Szalma and Taylor's (2011) work—based on significant empirical evidence, they write: "[T]he impact of automation and task demand on participant response depends to a substantial degree on human personality traits. A practical implication is that the identification of specific profiles to predict human–automation interaction is unlikely to be useful as a generic selection measure […] the complexity of the interactive effects suggest[s] that for selection tools to be effective, the optimal trait profiles would need to be developed separately for each level, type, and perhaps even domain of application of automation" (p. 91). We make a call for future studies on such interaction effects and their consequences for adoption decisions.

Design implications and adaptive systems as possible future design science research focus

Design science research has evolved as a major academic paradigm in the IS discipline, which aims to design innovative and useful IT artifacts (e.g., Hevner et al., 2004; March & Smith, 1995; Peffers et al., 2008). A major question in the current research context is how the presented empirical research can inform the design of AI systems in order to increase trust. In essence, though the 58 reviewed works have made valuable contributions to the literature (predominantly from a theoretical and empirical point of view, i.e., descriptive knowledge), the authors did not have an explicit intent to systematically conduct their research as design science research projects, nor was that the result. What follows is that *concrete* design implications (i.e., a form of prescriptive knowledge) of the current body of knowledge are limited.[26] This finding of the present review is consistent with a recent observation by Ahmad et al. (2022) who indicate that "previous studies mainly employ empirical methods to describe behaviors in interaction with these systems and consequently generate descriptive knowledge about the use and effectiveness of personality-adaptivity CAs [Conversational Agents, a specific type of AI system]. In addition […] there is still a research gap in providing prescriptive knowledge about how to design a PACA [personality-adaptive conversational agent] to improve interactions with users" (p. 3).

However, the analyzed literature offers some *abstract* design implications. Yet, these implications typically do not go beyond the recommendations which are already documented in existing conceptual papers or review works (e.g., Siau & Wang, 2018). Obviously, this constitutes a major limitation of the current body of knowledge which should be addressed in future research. As an example, in one of our 58 analyzed papers, Aliasghari et al. (2021) found, among other things, that "the participants lost trust in the robot […] when the robot made an error after it seemed to have learned the task by making no mistakes [… and] personality traits of the participants [...] affected some aspects of their trust in the robot" (p. 87). A design recommendation that would result from such an empirical finding is trivial: Developers should design AI systems which do not make errors. In a conceptual paper, Siau and Wang (2018) already indicated several years before the publication of the Aliasghari et al.'s study that trust "must be nurtured and maintained [… which] happens through […] competence of AI in completing tasks and finishing tasks in a consistent and reliable manner [… and] there should be no unexpected downtime or crashes" (p. 51).[27] Against the background of our finding that concrete design recommendations are hardly available in the current body of knowledge, we make a call for future IS design science research in the context of user personality and trust in AI systems. Then, more concrete design recommendations can be expected in the future.

As a starting point for future research, consider the following example within the mentioned domain of system errors. Foremost, it is *not* possible to create AI systems that do not make errors, independent from whether they are classical knowledge-based systems, are trained using machine learning algorithms, or are hybrids.[28] The aim must therefore be to prepare the user that errors might occur and how to cope with it (e.g., by implementing cross-check procedures or using an explanation component) (e.g., Shin, 2021). Also, expectation management is critical. Hence, it is important that the claims coming with an AI system are realistic because the gap between expectations and perceived delivery predominantly

[26] This paucity of design focus is substantiated by the fact that only one out of the 58 analyzed papers uses the term "design" in its title (Voinescu et al. 2018).

[27] Note that Siau & Wang (2018, pp. 50-52) describe several other factors which designers could consider in order to foster trust. Examples are representation (e.g., "a robot dog is another example of an AI representation that humans find easier to trust. Dogs are human's best friends and represent loyalty and diligence") or transparency and explainability (e.g., "challenges in machine learning and deep learning is the black box in the ML and decision-making processes. If the explainability of the AI application is poor or missing, trust is affected"), among several other factors (e.g., usability and reliability, collaboration and communication, sociability and bonding, security and privacy protection).

[28] The author would like to thank one anonymous reviewer for outlining this important point.

determines perceived quality and related trust perceptions (e.g., Jiang et al., 2002). However, today neither does scientific knowledge exist on how specific personality types perceive different preparation procedures, coping strategies, or expectation management approaches, nor do we know the procedures', strategies', and approaches' efficacy regarding trust as a function of user personality. This research gap should be closed in the future. Szalma and Taylor's (2011, p. 93) "general guidelines for incorporating individual differences into the design of human-technology interfaces" could serve as a starting point for corresponding design science research.

Moreover, a future design science research program should explicitly consider the development of adaptive systems. Recently, Matthews et al. (2021) substantiated the need for adaptive systems. They argued that task analysis is "a crucial step in the design of any system and its interfaces [… and] there should be an analogous 'person analysis'"—moreover, they foresee "individuation in design" and "adaptive technologies" because "conventional 'static' products offer little or no flexibility to accommodate variation in personality and individual differences once designed" (p. 7). Against this background, future studies should contribute to the conceptualization and development of such adaptive systems. Based on publications in the literature on real-time adaptive interventions (for an overview paper, see Nahum-Shani et al., 2018), we define such adaptive systems as a technology aiming to support the execution of a task by adapting to a user's personality, as well as to resulting neurophysiological states and behavior.

Imagine a situation in which a user interacts with an AI system. The AI system could be a chatbot or recommender system which is presented on the computer screen. In principle, however, the basic idea of adaptive systems also applies to other AI systems such as robots or autonomous vehicles. First, the system must be capable of recording the user state (e.g., emotions), its own state, context, and behavioral user data (e.g., preferences via system input in recommender systems). Second, the system must integrate this stream of data to derive the user's personality profile (either universal personality traits like the Big Five or specific personality traits like propensity to take risks or computer anxiety). Third, based on the personality profile the AI system adapts in real-time in

Table 4 Summary of example works in which AI systems adapt based on user personality

Study	Focus and major result
Ahmad et al. (2022)	The authors argue that currently conversational agents (CAs), a specific type of AI system, do not live up to their full potential because they are unable to provide responses tailored to users' personalities. To address this problem, the authors conducted a design science research project with the goal to design personality-adaptive conversational agents (PACAs). Based on an iterative and multi-step approach, they derive and formulate six design principles for PACAs for the domain of mental health care. The findings of the evaluation with psychologists and psychiatrists suggest that PACAs can be a promising source of mental health support. The authors indicate that with their design principles, they contribute to the body of design knowledge for CAs and provide guidance for practitioners who intend to design PACAs.
Kampman et al. (2019)	The authors describe a virtual agent whose major function is adaptation to user personality. User personality is determined through two approaches: audio data and speech transcription text. The virtual agent was implemented and evaluated. Based a Convolutional Neural Network, both approaches show good performance. The authors conclude that perception of user personality is possible using audio and text data. Moreover, the authors report correlations between user personality and preferred agent personality. Specifically, results suggest that the user personality openness correlates with a preference for agents with more gentle personality. Also, it was found that users report more empathy and enjoyed conversations more when agents adapted to their personality.
Tapus et al. (2008)	The authors describe a fully automated socially assistive therapist robot. The robot is able to monitor, assist, encourage, and socially interact with post-stroke patients in a rehabilitation context. The authors show that it is possible to develop a behavior adaptation system. Specifically, based on validation experiments it was demonstrated that the robot is able to adjust its social interaction parameters such as interaction distances, speed, or vocal content based on the user's personality and task performance. Moreover, it was demonstrated that adaptation to the user's personality may lead to improved human task performance.
Yorita et al. (2019)	Based on empirical evidence, the authors argue that people relate better with other people who have similar personality characteristics and that this fact is useful in scenarios where people should be receptive to receiving support and advice from others. The authors describe a chatbot personality model and a genetic algorithm that enables the chatbot to adapt its personality in real-time as it interacts with the user. The model draws upon the Big Five framework and focuses on extroversion and agreeableness. Results of the evaluation study indicate that the algorithms are capable of adapting personality traits to match the identified traits of the user.
Zalake (2020)	Interventions based on human interaction with virtual agents have been used to change user health-related attitudes and behaviors. The authors argue that current virtual agents use several persuasion techniques to influence user behavior. However, available technologies typically do not account for individual differences between users. This is problematic, because individual differences between users significantly affect the efficacy of different persuasion strategies. The study proposes a persuasive virtual agent-based mobile health intervention that adapts to individual differences in user personality to promote anxiety coping strategies.

order to maximize the user's trust in the system (foundations of such adaptive systems can be found, for example, in Byrne & Parasuraman, 1996, Harriott et al., 2018, or Picard, 1997).

It is critical to stress that research has already demonstrated the feasibility of adaptive systems based on user personality. Table 4 provides a summary of example works which could serve as a starting point for future design science research projects. Thus, despite the challenges which come along with the development of adaptive systems (e.g., Adam et al., 2017; Picard, 2003; vom Brocke et al., 2020), we foresee a prosperous development of research in this important domain. This positive outlook is substantiated by recent research which has demonstrated the feasibility of unobtrusive real-time measurement of trust in machines based on physiological measurement (Akash et al., 2018; Walker et al., 2019). Importantly, the development of such adaptive systems must not ignore the legal, ethical, organizational, and social aspects, which need to be explored in future studies. Because the development of adaptive systems is a highly interdisciplinary phenomenon, we expect a number of different scientific disciplines to contribute to this research (e.g., IS, informatics, organization science, and psychology). However, considering the IS discipline's successful history in interdisciplinary research endeavors, IS researchers could take a lead role in this area.

Limitations and concluding statement

In the present article, we reviewed 58 papers published in the period 2008-2021. To the best of our knowledge, this analysis is the first systematic review of the scientific literature on the relationship between personality traits and trust in the context of AI systems. Yet, the present review has limitations: First, while we can safely assume that the literature basis of the current review is extensive (because we searched for literature with several keywords and in four databases, including several hundred relevant journals and conference proceedings), there is no formal measure to prove completeness. Second, while we consider our literature collection and analysis approaches as rigorous, our interpretation of the findings, and especially the formulation of future research implications, is predominantly interpretive. Yet, because we report the analyzed literature in a very detailed way (see also Appendix 2), other researchers can directly draw upon our basis of 58 papers to complement or revise the current interpretations. What follows is that while we consider our review as a systematic documentation of the research status on personality and trust in the context of AI systems, it is hoped that this article instigates further research, thereby being more of a sort of "starting point" rather than the "final word". It will be rewarding to see what insights future research will reveal.

Appendix 1

The following list summarizes the 58 references which constitute the basis of the literature review.

1 Aliasghari, P., Ghafurian, M., Nehaniv, C. L., & Dautenhahn, K. (2021). Effect of domestic trainee robots' errors on human teachers' trust. *2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021*, 81–88. https://doi.org/10.1109/RO-MAN50785.2021.9515510.

2 Antes, A. L., Burrous, S., Sisk, B. A., Schuelke, M. J., Keune, J. D., & DuBois, J. M. (2021). Exploring perceptions of healthcare technologies enabled by artificial intelligence: an online, scenario-based survey. *BMC Medical Informatics and Decision Making, 21*(1), 221. https://doi.org/10.1186/s12911-021-01586-8.

3 Bawack, R. E., Wamba, S. F., & Carillo, K. D. A. (2021). Exploring the role of personality, trust, and privacy in customer experience performance during voice shopping: Evidence from SEM and fuzzy set qualitative comparative analysis. *International Journal of Information Management, 58*, 102309. https://doi.org/10.1016/j.ijinfomgt.2021.102309.

4 Böckle, M., Yeboah-Antwi, K., & Kouris, I. (2021). Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces. In D. H. & N. S. (Eds.), Artificial intelligence in HCI. HCII 2021. *Lecture Notes in Computer Science* (Vol. 12797, pp. 3–20). Springer. https://doi.org/10.1007/978-3-030-77772-2_1.

5 Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between trust attitudes toward automation, hofstede's cultural dimensions, and Big Five personality traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 841–845. https://doi.org/10.1177/1541931213601192.

6 Chiou, M., McCabe, F., Grigoriou, M., & Stolkin, R. (2021). Trust, shared understanding and locus of control in mixed-initiative robotic systems. *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 684–691. https://doi.org/10.1109/RO-MAN50785.2021.9515476.

7 Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction, 31*(10), 692–702. https://doi.org/10.1080/10447318.2015.1070549.

8 Cohen, J. F., & Sergay, S. D. (2011). An empirical study of health consumer beliefs, attitude and intentions toward the use of self-service kiosks. *17th Americas Conference on Information Systems 2011, AMCIS 2011 Proceedings - All Submissions, 46*, 403–412.

9 Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence, 298*, 103503. https://doi.org/10.1016/j.artint.2021.103503.

10 Cramer, H., Evers, V., Kemper, N., & Wielinga, B. (2008). Effects of autonomy, traffic conditions and driver personality traits on attitudes and trust towards in-vehicle agents. *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 477–482. https://doi.org/10.1109/WIIAT.2008.326.

11 Elson, J. S., Derrick, D., & Ligon, G. (2018). Examining trust and reliance in collaborations between humans and automated agents. Proceedings of the 51st Hawaii International Conference on System Sciences, 430–439. https://doi.org/10.24251/HICSS.2018.056.

12 Elson, J. S., Derrick, D., & Ligon, G. (2020). Trusting a humanoid robot: exploring personality and trusting effects in a human-robot partnership. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 543–552. https://doi.org/10.24251/HICSS.2020.067.

13 Ferronato, P., & Bashir, M. (2020a). An examination of dispositional trust in human and autonomous system interactions. In K. M. (Ed.), human-computer interaction. Human values and quality of life. HCII 2020. *Lecture Notes in Computer Science* (Vol. 12183, pp. 420–435). Springer. https://doi.org/10.1007/978-3-030-49065-2_30.

14 Ferronato, P., & Bashir, M. (2020b). Does the propensity to take risks influence human interactions with autonomous systems? In C. I., N. E., & A. T. (Eds.), advances in human factors in cybersecurity. AHFE 2020. *Advances in Intelligent Systems and Computing* (Vol. 1219, pp. 23–29). Springer. https://doi.org/10.1007/978-3-030-52581-1_4.

15 Gibson, A. M., Alarcon, G. M., Jessup, S. A., & Capiola, A. (2020). "Do you still trust me?" effects of personality on changes in trust during an experimental task with a human or robot partner. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 5099–5108.

16 Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior, 115*, 106607. https://doi.org/10.1016/j.chb.2020.106607.

17 Handrich, M. (2021). Alexa, you freak me out – identifying drivers of innovation resistance and adoption of intelligent personal assistants. *ICIS 2021 Proceedings, 11*, 1–17.

18 Hanna, N., & Richards, D. (2015). The influence of users' personality on the perception of intelligent virtual agents' personality and the trust within a collaborative context. In F. K. et Al. (Ed.), CARE-MFSC 2015, CCIS (Vol. 541, pp. 31–47). Springer. https://doi.org/10.1007/978-3-319-24804-2_3.

19 Haring, K. S., Matsumoto, Y., & Watanabe, K. (2013). How do people perceive and trust a lifelike robot. *Proceedings of the World Congress on Engineering and Computer Science 2013, 1*, 425–430.

20 Harriott, C. E., Garver, S., & Cunha, M. (2018). A motivation for co-adaptive human-robot interaction. In C. J (Ed.), advances in human factors in robots and unmanned systems, advances in intelligent systems and computing (Vol. 595, pp. 148–160). Springer. https://doi.org/10.1007/978-3-319-60384-1_15.

21 Hegner, S. M., Beldad, A. D., & Brunswick, G. J. (2019). In automatic we trust: investigating the impact of trust, control, personality characteristics, and extrinsic and intrinsic motivations on the acceptance of autonomous vehicles. *International Journal of Human–Computer Interaction, 35*(19), 1769–1780. https://doi.org/10.1080/10447318.2019.1572353.

22 Huang, H.-Y., & Bashir, M. (2017). Personal influences on dynamic trust formation in human-agent interaction. *Proceedings of the 5th International Conference on Human Agent Interaction*, 233–243. https://doi.org/10.1145/3125739.3125749.

23 Huang, H.-Y., Twidale, M., & Bashir, M. (2020). 'If you agree with me, do i trust you?': an examination of human-agent trust from a psychological perspective. In B. R. Bi Y. & K. S. (Eds.), intelligent systems and applications. IntelliSys 2019. *Advances in Intelligent Systems and Computing* (Vol. 1038, pp. 994–1013). Springer. https://doi.org/10.1007/978-3-030-29513-4_73.

24 Kim, K. J., Park, E., Sundar, S. S., & del Pobil, A. P. (2012). The effects of immersive tendency and need to belong on human-robot interaction. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '12*, 207–208. https://doi.org/10.1145/2157689.2157758.

25 Kim, W., Kim, N., Lyons, J. B., & Nam, C. S. (2020). Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Applied Ergonomics, 85*, 103056. https://doi.org/10.1016/j.apergo.2020.103056.

26 Kraus, J., Scholz, D., & Baumann, M. (2020a). What's driving me? exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors, 63*(6), 1076–1105. https://doi.org/10.1177/0018720820922653.

27 Kraus, J., Scholz, D., Messner, E.-M., Messner, M., & Baumann, M. (2020b). Scared to trust? – predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology, 10*, 2917. https://doi.org/10.3389/fpsyg.2019.02917.

28 Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science, 20*(4), 440–458. https://doi.org/10.1080/1463922X.2018.1491071.

29 Lyons, J. B., Nam, C. S., Jessup, S. A., Vo, T. Q., & Wynne, K. T. (2020). The role of individual differences as predictors of trust in autonomous security robots. *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 1–5. https://doi.org/10.1109/ICHMS49158.2020.9209544.

30 Matsui, T. (2021). Relationship between users' trust in robots and belief in paranormal entities. *Proceedings of the 9th International Conference on Human-Agent Interaction*, 252–256. https://doi.org/10.1145/3472307.3484666.

31 Matthews, G., Lin, J., Panganiban, A. R., & Long, M. D. (2020). Individual differences in trust in autonomous robots: implications for transparency. *IEEE Transactions on Human-Machine Systems, 50*(3), 234–244. https://doi.org/10.1109/THMS.2019.2947592.

32 McBride, M., Carter, L., & Ntuen, C. (2012). The impact of personality on nurses' bias towards automated decision aid acceptance. *International Journal of Information Systems and Change Management, 6*(2), 132–146. https://doi.org/10.1504/IJISCM.2012.051148.

33 Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: dispositional and history-based trust in human-automation interactions. *Human Factors, 50*(2), 194–210. https://doi.org/10.1518/001872008X288574.

34 Merritt, S. M., Heimbaugh, H., Lachapell, J., & Lee, D. (2013). I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534. https://doi.org/10.1177/0018720812465081.

35 Miller, L., Kraus, J., Babel, F., & Baumann, M. (2021). More than a feeling—interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety. *Frontiers in Psychology, 12*, 592711. https://doi.org/10.3389/fpsyg.2021.592711.

36 Müller, L., Mattke, J., Maier, C., Weitzel, T., & Graser, H. (2019). Chatbot acceptance: A latent profile analysis on individuals' trust in conversational agents. *SIGMIS-CPR 2019 - Proceedings of the 2019 Computers and People Research Conference*, 35–42. https://doi.org/10.1145/3322385.3322392.

37 Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: an experimental approach to human–technology interactions online. *Frontiers in Psychology, 11*, 568256. https://doi.org/10.3389/fpsyg.2020.568256.

38 Perelman, B. S., Evans, A. W., & Schaefer, K. E. (2017). Mental model consensus and shifts during navigation system-assisted route planning. Proceedings of the Human Factors and Ergonomics Society 2017 *Annual Meeting, 61*(1), 1183–1187. https://doi.org/10.1177/1541931213601779.

39 Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors, 57*(4), 545–556. https://doi.org/10.1177/0018720814564422.

40 Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn, 9*(1), 137–154. https://doi.org/10.1515/pjbr-2018-0010.

41 Rossi, S., Conti, D., Garramone, F., Santangelo, G., Staffa, M., Varrasi, S., & Di Nuovo, A. (2020). The role of personality factors and empathy in the acceptance and performance of a social robot for psychometric evaluations. *Robotics, 9*(2), 39. https://doi.org/10.3390/robotics9020039.

42 Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). *Would you trust a (faulty) robot? proceedings of the tenth annual ACM/IEEE international conference on Human-robot interaction*, 141–148. https://doi.org/10.1145/2696454.2696497.

43 Sarkar, S., Araiza-Illan, D., & Eder, K. (2017). *Effects of faults, experience, and personality on trust in a robot co-worker*. ArXiv Preprint, 1703.02335. http://arxiv.org/abs/1703.02335.

44 Schaefer, K. E., & Scribner, D. R. (2015). Individual differences, trust, and vehicle autonomy. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), 786–790. https://doi.org/10.1177/1541931215591242.

45 Schaefer, K. E., & Straub, E. R. (2016). Will passengers trust driverless vehicles? Removing the steering wheel and pedals. *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 159–165. https://doi.org/10.1109/COGSIMA.2016.7497804.

46 Schmidt, P., & Biessmann, F. (2020). Calibrating Human-AI collaboration: impact of risk, ambiguity and transparency on algorithmic bias. In A. H. et Al. (Ed.), machine learning and knowledge extraction. CD-MAKE 2020. *Lecture Notes in Computer Science* (Vol. 12279, pp. 431–449). Springer. https://doi.org/10.1007/978-3-030-57321-8_24.

47 Schrum, M. L., Neville, G., Johnson, M., Moorman, N., Paleja, R., Feigh, K. M., & Gombolay, M. C. (2021). Effects of social factors and team dynamics on adoption of collaborative robot autonomy. *ACM/IEEE International Conference on Human-Robot Interaction*, 149–157. https://doi.org/10.1145/3434073.3444649.

48 Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon, 6*(8), e04572. https://doi.org/10.1016/j.heliyon.2020.e04572.

49 Sorrentino, A., Mancioppi, G., Coviello, L., Cavallo, F., & Fiorini, L. (2021). Feasibility study on the role of personality, emotion, and engagement in socially assistive robotics: a cognitive assessment scenario. *Informatics, 8*(2), 23. https://doi.org/10.3390/informatics8020023.

50 Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied, 17*(2), 71–96. https://doi.org/10.1037/a0024170.

51 Tenhundfeld, N. L., de Visser, E. J., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2020). Trust and distrust of automated parking in a tesla model x. *Human Factors, 62*(2), 194–210. https://doi.org/10.1177/0018720819865412.

52 Tong, S. T., Corriero, E. F., Matheny, R. G., & Hancock, J. T. (2018). Online daters' willingness to use recommender technology for mate selection decisions. *IntRS Workshop, 2225*, 45–52.

53 Voinescu, A., Morgan, P. L., Alford, C., & Caleb-Solly, P. (2018). Investigating older adults' preferences for functions within a human-machine interface designed for fully autonomous vehicles. In Z. J. & S. G. (Eds.), human aspects of IT for the aged population. Applications in health, assistance, and entertainment. ITAP 2018. *Lecture Notes in Computer Science* (Vol. 10927, pp. 445–462). Springer. https://doi.org/10.1007/978-3-319-92037-5_32.

54 Xu, K. (2019). First encounter with robot Alpha: How individual differences interact with vocal and kinetic cues in users' social responses. *New Media & Society, 21*(11–12), 2522–2547. https://doi.org/10.1177/1461444819851479.

55 Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). *How do visual explanations foster end users' appropriate trust in machine learning? 25th International Conference on Intelligent User Interfaces (IUI '20)*, 189–201. https://doi.org/10.1145/3377325.3377480.

56 Youn, S., & Jin, S. V. (2021). "In A.I. we trust?" The effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging "feeling economy." *Computers in Human Behavior, 119*, 106721. https://doi.org/10.1016/j.chb.2021.106721.

57 Zhang, T., Tao, D., Qu, X., Zhang, X., Zeng, J., Zhu, H., & Zhu, H. (2020). Automated vehicle acceptance in China: Social influence and initial trust are key determinants. *Transportation Research Part C: Emerging Technologies, 112*, 220–233. https://doi.org/10.1016/j.trc.2020.01.027.

58 Zhou, J., Luo, S., & Chen, F. (2020). Effects of personality traits on user trust in human–machine collaborations. *Journal on Multimodal User Interfaces, 14*(4), 387–400. https://doi.org/10.1007/s12193-020-00329-9.

Appendix 2

The following table summarizes the major characteristics of the 58 reviewed papers and their underlying studies.

No.	Discipline	Universal traits	Specific traits	Context	Method	Sample size (N)	Mean age (years)	Female (in %)	Country (data collection)
1 P	CompSci	Big Five	Trust propensity	Domestic robot for food preparation	St1: Survey St2: Onl Exp	St1: 173 St2: 138	St1: 36.75 St2: 38.46	St1: 39% St2: n.a.	St1: Canada St2: Canada
2 J	CompSci	Big Five		AI-driven healthcare technologies	Survey	936	37.1	45%	USA
3 J	IS	Big Five		AI-based voice shopping	Survey	224	< 21 (6), 21–40 (149), 41–55 (52), 56–74 (16), > 74 (1)	44%	USA
4 P	HCI	Big Five		AI-enabled user interfaces	Survey	211	15–25 (4%), 26–35 (46%), 36–45 (28%), > 45 (22%)	41%	Different countries
5 P	Ergonom	Big Five		Air traffic control automation	Survey	360	20.92	n.a.	Different countries
6 P	CompSci		Locus of control	Human-robot interaction	Lab Exp	20	30.4	35%	UK
7 J	HCI		Locus of control	Autonomous vehicle	Survey	552	< 30 (31.9%), 30–39 (56.2%), 40–49 (7.6%), 50–59 (2.5%), > 59 (1.8%)	30.1%	Korea
8 P	IS		Computer anxiety, self-efficacy, need for interaction	Self-service kiosks in healthcare	Survey	192	18–25 (26), 26–40 (77), 41–55 (62), 56+ (20), missing (7)	56%	South Africa
9 J	CompSci	Big Five	Need for cognition	Intelligent tutoring system	Lab Exp	47	n.a.	79%	Canada
10 P	CompSci		Locus of control	In-vehicle agents	Onl Exp	100	33	26%	Different countries
11 P	IS	E		Automated recommendation agent	Lab Exp	64	23	52%	USA
12 P	IS	Big Five		Human-robot interaction	Lab study	58	21.69	n.a.	USA
13 P	HCI		Propensity to take risks	Autonomous system	Survey	344	23–38 (54.06%)	48.58%	USA
14 P	Ergonom	Big Five	Dispositional trust in automation, trust propensity	Autonomous system	Survey	344	23–38 (54.06%)	48.58%	USA

No.	Discipline	Universal traits	Specific traits	Context	Method	Sample size (N)	Mean age (years)	Female (in %)	Country (data collection)
15 P	IS		Trust propensity, distrust propensity, suspicion propensity	Interaction with an automated partner in a trust game context	Lab Exp	49	23.04	57%	USA
16 J	HCI		Attachment style	Artificial intelligence	St1: Survey St2+3: Onl Exp	St1: 248 St2: 374 St3: 272	St1: 18−80 St2: 18−78 St3: 18−82	St1: 72% St2: 61% St3: 65%	St1: USA St2: USA St3: USA
17 P	IS		Need for interaction, self-efficacy, technological innovativeness, novelty seeking	Intelligent Personal Assistants	Survey	168	32.7	46.4%	Germany
18 P	CompSci	A, E		Intelligent Virtual Agents	Lab Exp	55	22.56	n.a.	Australia
19 P	CompSci	Eysenck		Human-robot interaction (trust game)	Lab Exp	55	22.6	67%	Japan
20 P	Ergonom	O, N	Need for cognition, locus of control, trust propensity	Human-robot interaction	Lab Exp	28	18−55	61%	USA
21 J	HCI		Innovativeness, driving enjoyment	Autonomous vehicles	Survey	369	31	56%	Germany
22 P	HCI		Trust propensity, dispositional trust in automation	Automated security officer	Onl Exp	156	< 30 (43.6%), 30−60 (50.7%), > 60 (5.8%)	38%	Different countries
23 P	CompSci	A, E, N	Dispositional trust in automation, trust propensity	Automated agent (airport)	Onl Exp	156	< 30 (43.6%), 30−60 (50.7%), >60 (5.8%)	38%	Different countries
24 P	CompSci		Immersive tendency, need to belong,	Social robots	Lab Exp	20	22.75	60%	USA
25 J	Ergonom		Perfect Automation Schema (PAS)	Autonomous security robot	Onl Exp	233	33.75	46%	n.a.
26 J	Ergonom	Big Five	Self-esteem, locus of control, affinity for technology, self-efficacy, trust propensity, dispositional trust in automation	Advanced driver-assistance system	Survey	St1: 274 St2: 149	St1: 33.88 St2: 29.96	St1: 58% St2: 62%	St1: Germany St2: Germany
27 J	Psychol		Depressiveness, self-efficacy, self-esteem, locus of control	Automated driving system	Lab study	47	27.45	57%	Germany
28 J	Ergonom		Perfect Automation Schema (PAS)	St1+2: Human-robot interaction St3: Advanced automated safety system	St1: Lab Exp St2: Lab study St3: Survey	St1: 38 St2: 130 St3: 100	St1: 35 St2: 24 St3: n.a.	St1: 39% St2: 67% St3: n.a.	St1: USA St2: USA St3: USA

No.	Discipline	Universal traits	Specific traits	Context	Method	Sample size (N)	Mean age (years)	Female (in %)	Country (data collection)
29 P	CompSci	Big Five	Perfect Automation Schema (PAS)	Autonomous security robot	Survey	316	39	n.a.	n.a.
30 P	CompSci		Belief in paranormal entities	Human-robot interaction	Survey	59	44.6	24%	Japan
31 J	CompSci		Attitude towards robots	Human-robot interaction	Onl Exp	82	18–40	35%	n.a.
32 J	IS	Myers-Briggs		Automated decision aids (nursing)	Lab study	55	35	100%	USA
33 J	Ergonom	E	Dispositional trust in automation	Automatic weapon detector	Lab Exp	255	19.25	n.a.	USA
34 J	Ergonom		Implicit attitude towards automated systems, dispositional trust in automation	Automated baggage inspector	Onl Exp	69	25	77%	USA
35 J	Psychol		Dispositional trust in automation, attitude towards robots	Home assistance robot	Lab Exp	28	30.32	57%	Germany
36 P	IS	HEXACO		Chatbot	Survey	112	21–30 (44.6%), 31–40 (27.7%), 41–50 (12.5%), 51–60 (8%), > 60 (7.2%)	51.8%	Different countries
37 J	Psychol	Big Five	Robot use self-efficacy	Robot and AI support in decision-making trust game	Onl Exp	1,077	37.39	50.6%	USA
38 P	Ergonom		Mental model	Navigation technology	Lab Exp	27	35.67	41%	USA
39 J	Ergonom		Expectancy that automation is trustworthy	Baggage screening technology	Lab Exp	225	19.94	48%	USA
40 J	CompSci	Big Five	Trust propensity	Home companion robot	Onl Exp	200	33.56	43%	Different countries
41 J	CompSci	NEO-PI-3		Socially assistive robot	Lab Exp	19	61	42%	Italy
42 P	CompSci	E, N		Human-robot interaction	Lab Exp	40	37.95	55%	UK
43 *	arXiv	Big Five	Attitude towards robots	Human-robot interaction	Lab Exp	18	18–24 (12), 25–29 (4), 30–39 (2)	22%	UK
44 P	Ergonom		Dislike of driving, fatigue proneness, thrill seeking, driving aggression	Autonomous passenger vehicle	Lab Exp	24	n.a.	46%	USA
45 P	Ergonom		Intellect (having a vivid imagination)	Driverless vehicle	Lab Exp	20	18–64	30%	USA
46 P	CompSci		Affinity to risk	Machine learning support in decision-making	Onl Exp	248	n.a.	n.a.	n.a.

No.	Discipline	Universal traits	Specific traits	Context	Method	Sample size (N)	Mean age (years)	Female (in %)	Country (data collection)
47 P	CompSci		Propensity to trust	Human-robot interaction	Lab Exp	42	21.1	52.4%	USA
48 J	Other	Big Five	Locus of control	AI support in decision-making card game	Onl Exp	171	22.6	78%	Different countries
49 J	CompSci	Big Five		Socially assistive robotics	Lab study	8	83.33	38%	Italy
50 J	Psychol	Big Five		Simulated uninhabited ground vehicle	Lab Exp	161	19.8	32%	USA
51 J	Ergonom		Risk-taking behavior, self-confidence	Partially automated parking with a Tesla Model X	Field study	23	18–22	17%	USA
52 P	HCI		Need for cognition	Recommender technology	Lab Exp	129	20.35	76%	USA
53 P	HCI		Impulsive sensation seeking, aggression hostility, sociability, activity, neuroticism anxiety, attitude towardss computers	Autonomous Vehicle	Lab study	31	67.52	61.3%	UK
54 J	Other		Attitude towardss robots	Social robot	Lab Exp	110	20.44	50%	USA
55 P	HCI		Propensity to trust	Visual explanations for machine learning classifiers	Lab Exp	33	25–60	58%	USA
56 J	HCI		Ideological view (technopian vs. luddite)	Chatbot	Onl Exp	602	40.8	43%	USA
57 J	Other	Big Five	Sensation seeking	Automated Vehicle	Survey	604	31.38	42.9%	China
58 J	HCI	Big Five		Water pipe failure prediction technology	Lab Exp	42	30.4	24%	Australia

Notes: We harmonized terminology across the 58 papers in the column "Specific traits". In particular, we use the term "trust propensity" despite the fact that similar terms like "interpersonal propensity to trust", "disposition to trust", "dispositional trust", "dispositional interpersonal trust", "dispositional trust in humans", or "trust disposition" are used in the literature. Abbreviations: *J* Journal, *P* Proceedings, * = arXiv. *CompSci* Computer Science, Informatics, Robotics, *HCI* Human-Computer Interaction, *Ergonom* Ergonomics, Human Factors, *Psychol* Psychology. *IS* Information Systems, *A* Agreeableness, *E* Extraversion, *N* Neuroticism, *O* Openness, *Lab Exp* Laboratory Experiment, *Onl Exp* Online Experiment, *St* Study. *n.a.* not available

Funding Open access funding provided by University of Applied Sciences Upper Austria.

References

Adam, M. T. P., Gimpel, H., Maedche, A., & Riedl, R. (2017). Design blueprint for stress-sensitive adaptive Enterprise systems. *Business & Information Systems Engineering, 59*(4), 277–291. https://doi.org/10.1007/s12599-016-0451-3

Adorno, T., Frenkel-Brunswik, E., Levinson, D., & Sanford, N. (1950). *The authoritarian personality*. Harper.

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems, 8*(4), 1–20. https://doi.org/10.1145/3132743

Aliasghari, P., Ghafurian, M., Nehaniv, C. L., & Dautenhahn, K. (2021). Effect of domestic trainee robots' errors on human teachers' trust. In *2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021* (pp. 81–88). https://doi.org/10.1109/RO-MAN50785.2021.9515510

Antes, A. L., Burrous, S., Sisk, B. A., Schuelke, M. J., Keune, J. D., & DuBois, J. M. (2021). Exploring perceptions of healthcare technologies enabled by artificial intelligence: an online, scenario-based survey. *BMC Medical Informatics and Decision Making, 21*(1), 221. https://doi.org/10.1186/s12911-021-01586-8

Ashton, M. C., & Lee, K. (2016). Age trends in HEXACO-PI-R self-reports. *Journal of Research in Personality, 64*, 102–111. https://doi.org/10.1016/j.jrp.2016.08.008

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., Boies, K., & De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*(2), 356–366. https://doi.org/10.1037/0022-3514.86.2.356

Astor, P. J., Adam, M. T. P., Jerčić, P., Schaaff, K., & Weinhardt, C. (2013). Integrating biosignals into information systems: A NeuroIS tool for improving emotion regulation. *Journal of Management Information Systems, 30*(3), 247–278. https://doi.org/10.2753/MIS0742-1222300309

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bawack, R. E., Wamba, S. F., & Carillo, K. D. A. (2021). Exploring the role of personality, trust, and privacy in customer experience performance during voice shopping: Evidence from SEM and fuzzy set qualitative comparative analysis. *International Journal of Information Management, 58*, 102309. https://doi.org/10.1016/j.ijinfomgt.2021.102309

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artifical Intelligence. *MIS Quarterly, 45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Böckle, M., Yeboah-Antwi, K., & Kouris, I. (2021). Can you trust the black box? the effect of personality traits on trust in AI-enabled user interfaces. In Degen, H., & Ntoa, S. (Ed.), *Artificial Intelligence in HCI. HCII 2021. Lecture notes in computer science* (Vol. 12797, pp. 3–20). Springer. https://doi.org/10.1007/978-3-030-77772-2_1

Byrne, E. A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological Psychology, 42*(3), 249–268. https://doi.org/10.1016/0301-0511(95)05161-9

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *The handbook for the sixteen personality factor questionnaire*. Edited by the Institute for Personality and Ability Testing.

Chen, W., & Hirschheim, R. (2004). A paradigmatic and methodological examination of information systems research from 1991 to 2001. *Information Systems Journal, 14*(3), 197–235. https://doi.org/10.1111/j.1365-2575.2004.00173.x

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between Trust Attitudes Toward Automation, Hofstede's Cultural Dimensions, and Big Five Personality Traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 841–845. https://doi.org/10.1177/1541931213601192

Chiou, M., McCabe, F., Grigoriou, M., & Stolkin, R. (2021). Trust, shared understanding and locus of control in mixed-initiative robotic systems. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (pp. 684–691). https://doi.org/10.1109/RO-MAN50785.2021.9515476

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction, 31*(10), 692–702. https://doi.org/10.1080/10447318.2015.1070549

Cohen, J. F., & Sergay, S. D. (2011). An empirical study of health consumer beliefs, attitude and intentions toward the use of self-service kiosks. In *17th Americas Conference on Information Systems 2011, AMCIS 2011 Proceedings - All Submissions* (Vol. 46, pp. 403–412).

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management, 60*, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence, 298*, 103503. https://doi.org/10.1016/j.artint.2021.103503

Costa Jr., P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) manual*. Psychological Assessment Resources.

Cramer, H., Evers, V., Kemper, N., & Wielinga, B. (2008). Effects of autonomy, traffic conditions and driver personality traits on attitudes and trust towards in-vehicle agents. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008*, 477–482. https://doi.org/10.1109/WIIAT.2008.326

Dabholkar, P. A. (1992). The role of prior behavior and category-based affect in on-site service encounters. In J. F. Sherry & B. Sternthal (Eds.), *Diversity in consumer behavior: Vol. XIX (pp. 563–569)*. Association for Consumer Research.

Demazure, T., Karran, A., Léger, P.-M., Labonté-LeMoyne, É., Sénécal, S., Fredette, M., & Babin, G. (2021). Enhancing sustained attention. *Business & Information Systems Engineering, 63*(6), 653–668. https://doi.org/10.1007/s12599-021-00701-3

Devaraj, U. S., Easley, R. F., & Michael Crant, J. (2008). How does personality matter? Relating the five-factor model to technology acceptance and use. *Information Systems Research, 19*(1), 93–105. https://doi.org/10.1287/isre.1070.0153

Dimoka, A., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Gefen, D., Gupta, A., Ischebeck, A., Henning, P. H., Pavlou, P. A., Müller-Putz, G., Riedl, R., vom Brocke, J., & Weber, B. (2012). On the use of neurophysiological tools in IS research: Developing a research agenda for NeuroIS. *MIS Quarterly, 36*(3), 679–702. https://doi.org/10.2307/41703475

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y. Dwivedi. R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B.,Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-Fitz Hugh, L. C.,Misra, S., Mogaji, E., Kumar Sharma, S., Bahadur Singh, J., Raghavan, V., Ramanu, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., & Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*(101994), 1–47. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Elson, J. S., Derrick, D., & Ligon, G. (2018). Examining trust and reliance in collaborations between humans and automated agents. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 430–439). https://doi.org/10.24251/HICSS.2018.056

Elson, J. S., Derrick, D., & Ligon, G. (2020). Trusting a humanoid robot: exploring personality and trusting effects in a human-robot partnership. In *Proceedings of the 53rd Hawaii International*

Conference on System Sciences (pp. 543–552). https://doi.org/10.24251/HICSS.2020.067
Eysenck, H. J. (1947). *Dimensions of personality*. Kegan Paul.
Eysenck, H. J., & Eysenck, S. B. G. (1976). *Psychoticism as a dimension of personality*. Hodder & Stoughton.
Ferronato, P., & Bashir, M. (2020a). An examination of dispositional trust in human and autonomous system interactions. In K. M. (Ed.), *human-computer interaction. human values and quality of life. HCII 2020. Lecture notes in computer science* (Vol. 12183, pp. 420–435). Springer. https://doi.org/10.1007/978-3-030-49065-2_30
Ferronato, P., & Bashir, M. (2020b). Does the propensity to take risks influence human interactions with autonomous systems? In I. Corradini, E. Nardelli, & T. Ahram (Eds.), *Advances in human factors in cybersecurity. AHFE 2020. Advances in intelligent systems and computing* (Vol. 1219, pp. 23–29). Springer. https://doi.org/10.1007/978-3-030-52581-1_4
Funder, D. C. (2001). Personality. *Annual Review of Psychology, 52*(1), 197–221. https://doi.org/10.1146/annurev.psych.52.1.197
Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega, 28*(6), 725–737. https://doi.org/10.1016/S0305-0483(00)00021-9
Gibson, A. M., Alarcon, G. M., Jessup, S. A., & Capiola, A. (2020). "Do you still trust me?" effects of personality on changes in trust during an experimental task with a human or robot partner. In *Proceedings of the 53rd Hawaii International Conference on System Sciences* (pp. 5099–5108).
Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior, 115*, 106607. https://doi.org/10.1016/j.chb.2020.106607
Gist, M. E. (1987). Self-efficacy: Implications for organizational behavior and human resource management. *The Academy of Management Review, 12*(3), 472. https://doi.org/10.2307/258514
Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057
Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216
Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26
Hampson, S. E., Goldberg, L. R., & John, O. P. (1987). Category-breadth and social-desirability values for 573 personality terms. *European Journal of Personality, 1*(4), 241–258. https://doi.org/10.1002/per.2410010405
Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517–527. https://doi.org/10.1177/0018720811417254
Handrich, M. (2021). Alexa, you freak me out – Identifying drivers of innovation resistance and adoption of Intelligent Personal Assistants. *ICIS 2021 Proceedings, 11*, 1–17.
Hanna, N., & Richards, D. (2015). The influence of users' personality on the perception of intelligent virtual agents' personality and the trust within a collaborative context. In F. K. et al. (Eds.), *CARE-MFSC 2015* (Vol. 541, pp. 31–47). CCIS, Springer. https://doi.org/10.1007/978-3-319-24804-2_3
Haring, K. S., Matsumoto, Y., & Watanabe, K. (2013). How do people perceive and trust a lifelike robot. *Proceedings of the World Congress on Engineering and Computer Science, 2013*(1), 425–430.
Harriott, C. E., Garver, S., & Cunha, M. (2018). A motivation for co-adaptive human-robot interaction. In C. J. (Ed.), Advances in human factors in robots and unmanned systems, advances in intelligent systems and computing (Vol. 595, pp. 148–160). Springer. https://doi.org/10.1007/978-3-319-60384-1_15.
Harriott, C. E., Garver, S., & Cunha, M. (2018). A motivation for co-adaptive human-robot interaction. In J. Chen (Ed.), *Advances in human factors in robots and unmanned systems. AHFE 2017. Advances in intelligent systems and computing* (Vol. 595). Springer. https://doi.org/10.1007/978-3-319-60384-1_15
Harter, S. (1993). Causes and consequences of low self-esteem in children and adolescents. In R. Baumeister (Ed.), *Self-esteem: The puzzle of low self-regard* (pp. 87–116). Plenum. https://doi.org/10.1007/978-1-4684-8956-9_5
Hegner, S. M., Beldad, A. D., & Brunswick, G. J. (2019). In automatic we trust: investigating the impact of trust, control, personality characteristics, and extrinsic and intrinsic motivations on the acceptance of autonomous vehicles. *International Journal of Human–Computer Interaction, 35*(19), 1769–1780. https://doi.org/10.1080/10447318.2019.1572353
Hess, T., Fuller, M., & Campbell, D. (2009). Designing interfaces with social presence: Using vividness and extraversion to create social recommendation agents. *Journal of the Association for Information Systems, 10*(12), 889–919. https://doi.org/10.17705/1jais.00216
Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625
Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570
Hofstede, G., & McCrae, R. R. (2004). Personality and culture revisited: Linking traits and dimensions of culture. *Cross-Cultural Research, 38*(1), 52–88. https://doi.org/10.1177/1069397103259443
Huang, H.-Y., & Bashir, M. (2017). Personal influences on dynamic trust formation in human-agent interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction* (pp. 233–243). https://doi.org/10.1145/3125739.3125749
Huang, H.-Y., Twidale, M., & Bashir, M. (2020). 'If you agree with me, do I trust you?': An examination of human-agent trust from a psychological perspective. In Y. Bi & Y. Bi (Eds.), *Intelligent Systems and Applications. IntelliSys 2019. Advances in Intelligent Systems and Computing* (Vol. 1038, pp. 994–1013). Springer. https://doi.org/10.1007/978-3-030-29513-4_73
Jacovi, A., Marasovi, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency,* pp. 624–635. https://doi.org/10.1145/3442188.3445923.
Jiang, J. J., Klein, G., & Carr, C. L. (2002). Measuring information system service quality: SERVQUAL from the other side. *MIS Quarterly, 26*(2), 145–166 https://www.jstor.org/stable/4132324
John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research*. The Guilford Press.
John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). The Guilford Press.
Jung, C. G. (1923). *Psychological types*. Harcourt, Brace.
Kampman, O., Siddique, F. B., Yang, Y., & Fung, P. (2019). Adapting a virtual agent to user personality. In M. Eskenazi, L. Devillers, & J. Mariani (Eds.), *Advanced social interaction with agents. Lecture notes in electrical engineering* (Vol. 510, pp. 111–118). Springer. https://doi.org/10.1007/978-3-319-92108-2_13

Kim, K. J., Park, E., Sundar, S. S., & del Pobil, A. P. (2012). The effects of immersive tendency and need to belong on human-robot interaction. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '12* (pp. 207–208). https://doi.org/10.1145/2157689.2157758

Kim, W., Kim, N., Lyons, J. B., & Nam, C. S. (2020). Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Applied Ergonomics, 85*, 103056. https://doi.org/10.1016/j.apergo.2020.103056

Klein, H. K., & Hirschheim, R. (2008). The structure of the IS discipline reconsidered: Implications and reflections from a community of practice perspective. *Information and Organization, 18*(4), 280–302. https://doi.org/10.1016/j.infoandorg.2008.05.001

Kraus, J., Scholz, D., & Baumann, M. (2020a). What's driving me? Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors, 63*(6), 1076–1105. https://doi.org/10.1177/0018720820922653

Kraus, J., Scholz, D., Messner, E.-M., Messner, M., & Baumann, M. (2020b). Scared to trust? – Predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology, 10*, 2917. https://doi.org/10.3389/fpsyg.2019.02917

Lankton, N., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems, 16*(10), 880–918. https://doi.org/10.17705/1jais.00411

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Leung, A. K.-Y., & Cohen, D. (2011). Within- and between-culture variation: Individual differences and the cultural logics of honor, face, and dignity cultures. *Journal of Personality and Social Psychology, 100*(3), 507–526. https://doi.org/10.1037/a0022151

Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science, 20*(4), 440–458. https://doi.org/10.1080/1463922X.2018.1491071

Lyons, J. B., Nam, C. S., Jessup, S. A., Vo, T. Q., & Wynne, K. T. (2020). The role of individual differences as predictors of trust in autonomous security robots. *IEEE International Conference on Human-Machine Systems (ICHMS)*, (pp. 1–5). https://doi.org/10.1109/ICHMS49158.2020.9209544

Maier, C. (2012). Personality within information systems research: A literature analysis. *Proceedings of the European Conference on Information Systems,* 101.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Mason, R. O., McKenney, J. L., & Copeland, D. G. (1997). An historical method for MIS research: Steps and assumptions. *MIS Quarterly, 21*(3), 307. https://doi.org/10.2307/249499

Matsui, T. (2021). Relationship between users' trust in robots and belief in paranormal entities. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (pp. 252–256). https://doi.org/10.1145/3472307.3484666

Matthews, G., Lin, J., Panganiban, A. R., & Long, M. D. (2020). Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Transactions on Human-Machine Systems, 50*(3), 234–244. https://doi.org/10.1109/THMS.2019.2947592

Matthews, G., Hancock, P. A., Lin, J., Panganiban, A. R., Reinerman-Jones, L. E., Szalma, J. L., & Wohleber, R. W. (2021). Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Personality and Individual Differences, 169*, 109969. https://doi.org/10.1016/j.paid.2020.109969

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

McBride, M., Carter, L., & Ntuen, C. (2012). The impact of personality on nurses' bias towards automated decision aid acceptance. *International Journal of Information Systems and Change Management, 6*(2), 132–146. https://doi.org/10.1504/IJISCM.2012.051148

McCarthy, J. L., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence*. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McCrae, R. R. (2000). Trait psychology and the revival of personality and culture studies. *American Behavioral Scientist, 44*(1), 10–31. https://doi.org/10.1177/00027640021956062

McCrae, R. R. (2004). Human nature and culture: A trait perspective. *Journal of Research in Personality, 38*(1), 3–14. https://doi.org/10.1016/j.jrp.2003.09.009

McCrae, R. R., & Costa, P. T. J. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

McCrae, R. R., & Costa, P. T. J. (1997). Personality trait structure as a human universal. *American Psychologist, 52*(5), 509–516. https://doi.org/10.1037/0003-066X.52.5.509

McCrae, R. R., & Costa, P. T. J. (1999). A Five-Factor Theory of personality. In A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). The Guilford Press.

McCrae, R. R., Costa, P. T. J., & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment, 84*(3), 261–270. https://doi.org/10.1207/s15327752jpa8403_05

McElroy, J. C., Hendrickson, A. R., Townsend, A. M., & DeMarie, S. M. (2007). Dispositional factors in internet use: Personality versus cognitive style. *MIS Quarterly, 31*(4), 809–820. https://doi.org/10.2307/25148821

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review, 23*(3), 473–490. https://doi.org/10.5465/amr.1998.926622

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: dispositional and history-based trust in human-automation interactions. *Human Factors, 50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Merritt, S. M., Heimbaugh, H., Lachapell, J., & Lee, D. (2013). I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors, 57*(5), 740–753. https://doi.org/10.1177/0018720815581247

Miller, L., Kraus, J., Babel, F., & Baumann, M. (2021). More than a feeling—interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety. *Frontiers in Psychology, 12*, 592711. https://doi.org/10.3389/fpsyg.2021.592711

Montag, C., & Panksepp, J. (2017). Primary emotional systems and personality: An evolutionary perspective. *Frontiers in Psychology, 8*, 464. https://doi.org/10.3389/fpsyg.2017.00464

Montag, C., Hahn, E., Reuter, M., Spinath, F. M., Davis, K., & Panksepp, J. (2016). The role of nature and nurture for individual differences in primary emotional systems: Evidence from a twin study. *PLoS One, 11*(3), e0151405. https://doi.org/10.1371/journal.pone.0151405

Montag, C., Elhai, J. D., & Davis, K. L. (2021). A comprehensive review of studies using the affective neuroscience personality scales in the psychological and psychiatric sciences. *Neuroscience & Biobehavioral Reviews, 125*, 160–167. https://doi.org/10.1016/j.neubiorev.2021.02.019

Mooradian, T., Renzl, B., & Matzler, K. (2006). Who trusts? Personality, trust and knowledge sharing. *Management Learning, 37*(4), 523–540. https://doi.org/10.1177/1350507606073424

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the Big Five personality traits and the big six vocational interest types. *Personnel Psychology, 58*(2), 447–478. https://doi.org/10.1111/j.1744-6570.2005.00468.x

Mühl, K., Strauch, C., Grabmaier, C., Reithinger, S., Huckauf, A., & Baumann, M. (2020). Get ready for being chauffeured: Passenger's preferences and trust while being driven by human and automation. *Human Factors, 62*(8), 1322–1338. https://doi.org/10.1177/0018720819872893

Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *The MBTI® manual: A guide to the development and use of the Myers-Briggs type indicator*. Consulting Psychologists Press.

Müller, L., Mattke, J., Maier, C., Weitzel, T., & Graser, H. (2019). Chatbot acceptance: A latent profile analysis on individuals' trust in conversational agents. In *SIGMIS-CPR 2019 - Proceedings of the 2019 Computers and People Research Conference* (pp. 35–42). https://doi.org/10.1145/3322385.3322392

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collings, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine, 52*(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8

Nam, T. (2019). Citizen attitudes about job replacement by robotic automation. *Futures, 109*, 39–49. https://doi.org/10.1016/j.futures.2019.04.005

Nilsson, N. J. (2010). *The quest for artificial intelligence: A history of ideas and achievements*. Cambridge University Press.

Orri, M., Pingault, J.-B., Rouquette, A., Lalanne, C., Falissard, B., Herba, C., Côté, S. M., & Berthoz, S. (2017). Identifying affective personality profiles: A latent profile analysis of the affective neuroscience personality scales. *Scientific Reports, 7*(1), 4548. https://doi.org/10.1038/s41598-017-04738-x

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology, 11*, 568256. https://doi.org/10.3389/fpsyg.2020.568256

Panganiban, A. R., Matthews, G., Lin, J., & Long, M. D. (2020). Trust your robot! Individual differences in confidence in robot threat evaluations. Abstracts from the International Society for the Study of individual differences conference 2019. *Personality and Individual Differences*, *157*(109684), 36.

Paravastu, N., Gefen, D., & Creason, S. B. (2014). Understanding trust in IT artifacts. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems, 45*(4), 30–50. https://doi.org/10.1145/2691517.2691520.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Perelman, B. S., Evans, A. W., & Schaefer, K. E. (2017). Mental model consensus and shifts during navigation system-assisted route planning. *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting, 61*(1), 1183–1187. https://doi.org/10.1177/1541931213601779

Picard, R. W. (1997). *Affective computing*. MIT Press.

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies, 59*(1–2), 55–64. https://doi.org/10.1016/S1071-5819(03)00052-1

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors, 57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Reuter, M., Panksepp, J., Davis, K. L., & Montag, C. (2017). *Affective neuroscience personality scales (ANPS) – Deutsche Version*. Hogrefe.

Riedl, R. (2021). Trust and Digitalization: Review of Behavioral and Neuroscience Evidence. In F. Krueger (Ed.), *The Neurobiology of Trust* (pp. 54-76). Cambridge University Press. https://doi.org/10.1017/9781108770880.005

Riedl, R., & Léger, P.-M. (2016). *Fundamentals of NeuroIS – Information systems and the brain*. Springer.

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn, 9*(1), 137–154. https://doi.org/10.1515/pjbr-2018-0010

Rossi, S., Conti, D., Garramone, F., Santangelo, G., Staffa, M., Varrasi, S., & Di Nuovo, A. (2020). The role of personality factors and empathy in the acceptance and performance of a social robot for psychometric evaluations. *Robotics, 9*(2), 39. https://doi.org/10.3390/robotics9020039

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80*(1), 1–28. https://doi.org/10.1037/h0092976

Rotter, J. B. (1990). Internal versus external control of reinforcement: A case history of a variable. *American Psychologist, 45*(4), 489–493. https://doi.org/10.1037/0003-066X.45.4.489

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot? *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, 141–148*. https://doi.org/10.1145/2696454.2696497

Sarkar, S., Araiza-Illan, D., & Eder, K. (2017). Effects of faults, experience, and personality on trust in a robot co-worker. *ArXiv Preprint, 1703*, 02335 http://arxiv.org/abs/1703.02335

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schmidt, P., & Biessmann, F. (2020). Calibrating human-AI collaboration: impact of risk, ambiguity and transparency on algorithmic bias. In A. H. et al. (Eds.), *Machine learning and knowledge extraction. CD-MAKE 2020. Lecture notes in computer science* (Vol. 12279, pp. 431–449). Springer. https://doi.org/10.1007/978-3-030-57321-8_24

Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon, 6*(8), e04572. https://doi.org/10.1016/j.heliyon.2020.e04572

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, 1–10. https://doi.org/10.1016/j.ijhcs.2020.102551

Schaefer, K. E., & Scribner, D. R. (2015). Individual differences, trust, and vehicle autonomy. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), 786–790. https://doi.org/10.1177/1541931215591242

Schaefer, K. E., & Straub, E. R. (2016). Will passengers trust driverless vehicles? Removing the steering wheel and pedals. *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2016*, 159–165. https://doi.org/10.1109/COGSIMA.2016.7497804

Schrum, M. L., Neville, G., Johnson, M., Moorman, N., Paleja, R., Feigh, K. M., & Gombolay, M. C. (2021). Effects of social factors and team dynamics on adoption of collaborative robot autonomy. *ACM/IEEE International Conference on Human-Robot Interaction, 149–157*. https://doi.org/10.1145/3434073.3444649

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*(2), 47–53.

Sindermann, C., Riedl, R., & Montag, C. (2020). Investigating the relationship between personality and technology acceptance with a focus on the smartphone from a gender perspective: Results of an exploratory survey study. *Future Internet, 12*(7), 110. https://doi.org/10.3390/fi1207011

Söllner, M., Hoffmann, A., Hoffmann, H., Wacker, A., & Leimeister, J. M. (2012). Understanding the formation of trust in IT artifacts. *ICIS 2012 Proceedings*. 11. https://aisel.aisnet.org/icis2012/proceedings/HumanBehavior/11

Sorrentino, A., Mancioppi, G., Coviello, L., Cavallo, F., & Fiorini, L. (2021). Feasibility study on the role of personality, emotion, and engagement in socially assistive robotics: A cognitive assessment scenario. *Informatics, 8*(2), 23. https://doi.org/10.3390/informatics8020023

Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior, 59*(2), 364–382. https://doi.org/10.1016/j.geb.2006.07.006

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied, 17*(2), 71–96. https://doi.org/10.1037/a0024170

Tapus, A., Ţăpuş, C., & Matarić, M. J. (2008). User—Robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics, 1*, 169–183. https://doi.org/10.1007/s11370-008-0017-4

Tenhundfeld, N. L., de Visser, E. J., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2020). Trust and distrust of automated parking in a tesla model X. *Human Factors, 62*(2), 194–210. https://doi.org/10.1177/0018720819865412

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Tong, S. T., Corriero, E. F., Matheny, R. G., & Hancock, J. T. (2018). Online daters' willingness to use recommender technology for mate selection decisions. *IntRS Workshop, 2225*, 45–52.

Tupes, E. C., & Christal, R. E. (1961). Recurrent personality factors based on trait ratings. USAF ASD tech. *Rep., 60*(61–97), 225–251. https://doi.org/10.1111/j.1467-6494.1992.tb00973.x

vom Brocke, J., Simons, A., Niehaves, B., Reimer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *Proceedings of the European Conference on Information Systems,* 2206–2217.

vom Brocke, J., Hevner, A., Léger, P. M., Walla, P., & Riedl, R. (2020). Advancing a NeuroIS research agenda with four areas of societal contributions. *European Journal of Information Systems, 29*(1), 9–24. https://doi.org/10.1080/0960085X.2019.1708218

Voinescu, A., Morgan, P. L., Alford, C., & Caleb-Solly, P. (2018). Investigating older adults' preferences for functions within a human-machine interface designed for fully autonomous vehicles. In Zhou, J., & Salvendy, G. (Ed.), *Human aspects of IT for the aged population. applications in health, assistance, and entertainment. ITAP 2018. Lecture notes in computer science* (Vol. 10927, pp. 445–462). Springer. https://doi.org/10.1007/978-3-319-92037-5_32

Walker, F., Wang, J., Martens, M. H., & Verwey, W. B. (2019). Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour, 64*, 401–412. https://doi.org/10.1016/j.trf.2019.05.021

Walsham, G. (1995). The emergence of interpretivism in IS research. *Information Systems Research, 6*(4), 376–394. https://doi.org/10.1287/isre.6.4.376

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2), xiii–xxiii.

Xu, K. (2019). First encounter with robot Alpha: How individual differences interact with vocal and kinetic cues in users' social responses. *New Media & Society, 21*(11–12), 2522–2547. https://doi.org/10.1177/1461444819851479

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *25th International Conference on Intelligent User Interfaces (IUI '20)* (pp. 189–201). https://doi.org/10.1145/3377325.3377480

Yorita, A., Egerton, S., Oakman, J., Chan, C., & Kubota, N. (2019). Self-adapting chatbot personalities for better peer support. *IEEE international conference on systems, man and cybernetics* (pp. 4094–4100) https://ieeexplore.ieee.org/document/8914583

Youn, S., & Jin, S. V. (2021). "In A.I. we trust?" The effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging "feeling economy". *Computers in Human Behavior, 119*, 106721. https://doi.org/10.1016/j.chb.2021.106721

Zalake, M. (2020). Advisor: Agent-based intervention leveraging individual differences to support mental wellbeing of college students. *CHI EA '20: Extended abstracts of the 2020 CHI conference on human factors in computing systems,* 1–8. https://doi.org/10.1145/3334480.3375026.

Zhang, P. (2013). The affective response model: A theoretical framework of affective concepts and their relationships in the ICT context. *MIS Quarterly, 37*(1), 247–274 https://www.jstor.org/stable/43825945

Zhang, T., Tao, D., Qu, X., Zhang, X., Zeng, J., Zhu, H., & Zhu, H. (2020). Automated vehicle acceptance in China: Social influence and initial trust are key determinants. *Transportation Research Part C: Emerging Technologies, 112*, 220–233. https://doi.org/10.1016/j.trc.2020.01.027

Zhou, J., Luo, S., & Chen, F. (2020). Effects of personality traits on user trust in human–machine collaborations. *Journal on Multimodal User Interfaces, 14*(4), 387–400. https://doi.org/10.1007/s12193-020-00329-9

Zimbardo, P., Johnson, R., & McCann, V. (2021). *Psychology: Core concepts (8^{th} edition)*. Pearson.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 12 August 2024
DOI 10.3389/fpsyg.2024.1422177

OPEN ACCESS

EDITED BY
Casey Bennett,
DePaul University, United States

REVIEWED BY
Georg Starke,
Swiss Federal Institute of Technology Lausanne, Switzerland
Nejc Plohl,
University of Maribor, Slovenia

*CORRESPONDENCE
René Riedl
✉ rene.riedl@fh-steyr.at

RECEIVED 23 April 2024
ACCEPTED 18 July 2024
PUBLISHED 12 August 2024

CITATION
Riedl R, Hogeterp SA and Reuter M (2024) Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice.
Front. Psychol. 15:1422177.
doi: 10.3389/fpsyg.2024.1422177

Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice

René Riedl[1,2]*, Svea A. Hogeterp[3] and Martin Reuter[3]

[1]Digital Business Institute, University of Applied Sciences Upper Austria, Steyr, Austria, [2]Institute of Business Informatics – Information Engineering, University of Linz, Linz, Austria, [3]Institute of Psychology, University of Bonn, Bonn, Germany

Today the doctor-patient relationship typically takes place in a face-to-face setting. However, with the advent of artificial intelligence (AI) systems, two further interaction scenarios are possible: an AI system supports the doctor's decision regarding diagnosis and/or treatment while interacting with the patient, or an AI system could even substitute the doctor and hence a patient interacts with a chatbot (i.e., a machine) alone. Against this background, we report on an online experiment in which we analyzed data from $N = 1{,}183$ people. The data was collected in German-speaking countries (Germany, Austria, Switzerland). The participants were asked to imagine they had been suffering from medical conditions of unknown origin for some time and that they were therefore visiting a health center to seek advice from a doctor. We developed descriptions of patient-doctor interactions (referred to as vignettes), thereby manipulating the patient's interaction partner: (i) human doctor, (ii) human doctor with an AI system, and (iii) an AI system only (i.e., chatbot). Furthermore, we manipulated medical discipline: (i) cardiology, (ii) orthopedics, (iii) dermatology, and (iv) psychiatry. Based on this 3×4 experimental within-subjects design, our results indicate that people prefer a human doctor, followed by a human doctor with an AI system, and an AI system alone came in last place. Specifically, based on these 12 hypothetical interaction situations, we found a significant main effect of a patient's interaction partner on trust, distrust, perceived privacy invasion, information disclosure, treatment adherence, and satisfaction. Moreover, perceptions of trust, distrust, and privacy invasion predicted information disclosure, treatment adherence, and satisfaction as a function of interaction partner and medical discipline. We found that the situation in psychiatry is different from the other three disciplines. Specifically, the six outcome variables differed strongly between psychiatry and the three other disciplines in the "human doctor with an AI system" condition, while this effect was not that strong in the other conditions (human doctor, chatbot). These findings have important implications for the use of AI in medical care and in the interaction between patients and their doctors.

KEYWORDS

artificial intelligence, chatbot, trust, doctor-patient relationship, medicine, patient-doctor interaction, privacy invasion, psychiatry

1 Introduction

The *relationship* between doctor and patient has been crucial since the early days of ancient civilizations, because it affects major medical outcomes such as patient information disclosure, treatment adherence, satisfaction, and ultimately treatment success (e.g., McGrew, 1985; Stewart, 1995; Elgood, 2010; Spikins et al., 2018). What has characterized the doctor-patient relationship in the past and still does today is that the interaction typically takes place face-to-face. However, due to technological advancements, changes in this mode of interaction have become evident. Telemedicine (defined as distribution of health-related services and information via information and communication technologies, ICT) has become, at least in some contexts, an accepted practice (e.g., Hollander and Carr, 2020; Mann et al., 2020). Example contexts are the provision of medical services in rural settings or if social distancing is required such as in the case of infectious diseases like COVID-19. Telemedicine, however, constitutes computer-mediated communication. Thus, ICT enables direct communication between a doctor and patient who are not physically in the same place, typically via videoconferencing tools. However, the patient still interacts with a *human* doctor.

Moreover, artificial intelligence (AI) systems have been discussed, as well as implemented, in various medical contexts such as automatic disease diagnosis or surgeries by robots (e.g., Hashimoto et al., 2018; Kelly et al., 2019). In line with Rai et al. (2019), we define AI "as the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity" (p. iii). Importantly, the use of AI systems also spans situations in which a patient interacts with a doctor. In addition to the traditional face-to-face doctor-patient interaction, two AI-based interaction styles have been discussed in the scientific literature and could constitute possible interaction scenarios in the future. First, an AI system *complements* the doctor's decision regarding diagnosis and/or treatment. Here, the AI system makes a recommendation to the doctor, who verifies the suggestion, overrules it if deemed necessary, and makes the final decision (e.g., Jussupow et al., 2021). Thus, such an AI system can be considered as a decision support system which is used when a doctor interacts with a patient. Second, an AI system *substitutes* the doctor. Here, the patient interacts with an "intelligent" system, but not with a human doctor (e.g., Yokoi et al., 2021). Interaction with such an "intelligent" system typically manifests as patient-chatbot interaction. In addition to computing power and an AI component, such a chatbot has a user interface, which is either text-based (i.e., the patient types on a keyboard and the AI system responds via the screen; comparable to ChatGPT) or speech-based (spoken communication instead of written, comparable to speech assistants like Amazon's Alexa or Apple's Siri) (e.g., Powell, 2019). Thus, what we can observe today are different degrees of AI involvement in doctor-patient interaction, with no involvement in classic human-human interaction, some involvement in situations where the doctor is assisted by the AI system, and maximum involvement in situations where the patient interacts only with a machine (i.e., chatbot).

In this article, we report on an online experiment in which participants were asked to imagine they had been suffering from medical conditions of unknown origin for some time and that they were therefore visiting a health center to seek advice from a doctor. We manipulated the patient's interaction partner—it was either a (i) human doctor, (ii) human doctor with AI system, or (iii) AI system only (i.e., a chatbot). Moreover, we also considered different medical disciplines: (i) cardiology, (ii) orthopedics, (iii) dermatology, and (iv) psychiatry to identify a possible interaction effect of the patient's interaction partner by medical discipline. The reason for this nuanced perspective is that medical disciplines function differently in terms of the nature of doctor-patient interaction. Psychiatric care, in particular, heavily relies on direct, empathetic human interactions, as these interactions are crucial for the diagnosis and treatment of psychiatric disorders, which are often stigmatized and require nuanced understanding and empathy from the caregiver (e.g., Starke et al., 2021, 2023). Patients in psychiatric settings may perceive a doctor's use of AI as a distraction, potentially reducing their sense of being cared for and understood, which are vital components of effective psychiatric treatment. In contrast, other medical disciplines such as cardiology, orthopedics, and dermatology involve conditions more directly related to body structures and hence may be less dependent on the emotional and empathetic exchanges critical in psychiatry. AI systems in these fields can assist in diagnosing and treating conditions where objective data and clear physiological markers are paramount, possibly improving efficiency without significantly impacting patient trust, satisfaction, or other important factors in doctor-patient interaction (e.g., Powell, 2019; Nagy and Sisk, 2020). Furthermore, AI's limitations in interpreting the complex and often subjective nature of psychiatric symptoms highlight the need for cautious integration of AI in psychiatric care. The inability of AI to fully grasp the intricacies of human emotions and social cues can lead to patient distrust and dissatisfaction, especially when privacy concerns and the potential for data breaches are at play (e.g., Asan et al., 2020). Given these differences, it is imperative to consider the specific requirements and patient perceptions in various medical disciplines when integrating AI into doctor-patient interaction.

Against this background, in the present study the participants were presented with 12 (3×4) lifelike descriptions of patient-doctor interactions and were asked to respond to questions after being confronted with each situation. The questions referred to six variables: trust, distrust, privacy invasion, information disclosure, treatment adherence, and satisfaction. We chose these six variables as they constitute major factors in patient-doctor interaction (Pearson and Raeke, 2000; Dalton-Brown, 2020; Nagy and Sisk, 2020; Wu et al., 2022).

One seminal conceptualization of trust defines interpersonal trust as the willingness of one party (i.e., the trustor) to be vulnerable to another (i.e., the trustee) based on the belief that the trustee will perform actions that are important to the trustor (Mayer et al., 1995). It follows that the patient is the trustor and the doctor is the trustee in the context of physician-patient interaction. More specifically, in this relationship trust includes the patient's confidence in the physician's competence, integrity, and commitment to act in the patient's best interest (e.g., Birkhäuer et al., 2017). Moreover, it includes the belief that the physician is knowledgeable, skilled, and able to provide effective care while being honest and respectful (e.g., Pearson and Raeke, 2000). Trust is also built through the physician's ability to listen to and address the patient's concerns and preferences and especially through confidentiality; therefore, trust enables patients to feel comfortable sharing personal information and following medical advice, ultimately contributing to better health outcomes and patient satisfaction (Pearson and Raeke, 2000; Birkhäuer et al., 2017; Powell, 2019; Gille et al., 2020). Notably, while the importance of trust has been recognized by scholars studying the doctor-patient relationship, the importance of distrust has not (e.g., Birkhäuer et al., 2017). A main reason for this lack of research is that scholars often consider trust and distrust as the opposite ends of a single construct continuum (e.g., Barber, 1983). In contrast, more recent behavioral research, both theoretical (e.g., Lewicki et al., 1998) and empirical (e.g., McKnight and Choudhury, 2006), identifies trust and distrust as separate constructs. Brain research has even found that trust and distrust perceptions activate different brain areas, which is considered the most direct evidence for the fact that trust and distrust are indeed separate constructs, despite the fact that self-report data on trust and distrust may be negatively correlated (Dimoka, 2010; Riedl and Javor, 2012; Krueger and Meyer-Lindenberg, 2019).

Moreover, to control for possible confounding effects, we measured some health status variables, demographics (e.g., sex, age), and several individual difference variables (e.g., personality, technology attitude) of our participants. Controlling for these variables is important because trust and distrust, as well as the other variables which are important in the doctor-patient interaction (i.e., privacy invasion, information disclosure, treatment adherence, and satisfaction), have been shown to correlate with sex (e.g., Riedl et al., 2010), age (e.g., Sutter and Kocher, 2007), personality (e.g., Bruce et al., 2010), and technology attitude (e.g., Nam, 2019), among others.

To the best of our knowledge, no published research has investigated individuals' preference structure[1] regarding the 12 manipulations (3 interaction partners × 4 medical disciplines) based on the variables trust, distrust, privacy invasion, information disclosure, treatment adherence, and satisfaction. Therefore, we investigated the following research question: *Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline?*

The present study contributes to the academic literature in a novel way, and—considering the sharp increase of AI systems in medical contexts in the recent past—also deals with a topic of significant practical relevance. The present study's contribution is further substantiated by conflicting perspectives and findings in the academic literature. On the one hand, the literature provides sound arguments why patients could prefer AI systems to human doctors. As an example, Dalton-Brown (2020) writes: "A patient may trust an AI more, believing the AI's health advice is unbiased and not prone to human fallibility. An AI doctor will not be exhausted after a long shift, nor has it any concept of power" (p. 118). On the other hand, recent evidence suggests that an AI system is trusted less than a human doctor in medical treatment decisions (Yokoi et al., 2021). In addition to these conflicting positions regarding patient preferences for a human doctor versus an AI system, we are unaware of empirical evidence on the effects of a blend of both (i.e., a doctor with an AI system). This research deficit is remarkable, as the literature on AI-assisted decision-making distinguishes direct from indirect user interaction with an AI system. In a recent review paper, Vereschak et al. (2021) report that 94% of the extant literature on AI-assisted decision-making examined direct interaction. Thus, also studying indirect interaction is an important research endeavor. Specifically, a patient does not interact directly with the AI system, but with a doctor who is supported by an AI system, and hence from the patient's perspective the interaction with the system is indirect.

1 A preference refers to a person's attitude towards a specific stimulus.

2 Materials and methods

2.1 Research approach, data collection, and sample

This study conforms to the Declaration of Helsinki.[2] All participants gave informed consent. We conducted an online experiment with vignettes based on a within-subjects design. Vignettes are "stories about individuals and situations which make reference to important points in the study of perceptions, beliefs, and attitudes" (Hughes, 1998, p. 381). Thus, vignettes are "simulations of real events" (Robert et al., 2009, p. 251). Vignettes have been applied successfully to study medical phenomena (e.g., Gould, 1996; Hughes, 1998).

In the current study, participants were presented 12 different vignettes: 3 patient's interaction partner (human doctor, human doctor with AI system, AI system alone) × 4 medical discipline (cardiology, orthopedics, dermatology, psychiatry) scenarios in which a patient-doctor interaction was described.[3] The vignettes are shown in Appendix A (note that all appendices to this paper are part of the Supplementary material). When developing the vignettes, care was taken to ensure that they differed only in the two manipulated variables (patient's interaction partner, medical discipline). The rest of the text was identical except for linguistic nuances (especially the description of symptoms was kept constant in each medical discipline across the three interaction partner conditions). After the first draft of the vignettes had been developed by the first author, these drafts were checked and revised with the other two co-authors for content and linguistic precision. In addition, the final set of vignettes developed by the authors was also discussed with five practitioners, which led to minor linguistic adjustments (e.g., the term "tachycardia" in the vignettes was replaced by the commonly understood description "a racing heart"). Participants were asked to put themselves in the patient role when answering the questions on trust, distrust, perceived privacy invasion, information disclosure, treatment adherence, and satisfaction. These questions were presented directly after each vignette. Thus, participants responded to the questions regarding these six constructs twelve times.

2 In our purely behavioral research study, the key ethical considerations from the Declaration of Helsinki include obtaining informed consent, ensuring beneficence and non-maleficence, maintaining confidentiality, conducting risk–benefit assessments, ensuring scientific validity, and ensuring credible, meaningful results.

3 These four medical disciplines were selected because they differ significantly. Importantly, in the preparation of the present study the authors considered more disciplines (e.g., neurology). However, based on feedback in the pre-testing phase it turned out that 12 different vignettes are the reasonable maximum regarding effort to complete the study. Note that any other discipline would have increased the number of vignettes by an additional 3.

Moreover, we collected data on the health status[4] and demographics (sex, age, educational attainment) before the presentation of the vignettes. The vignettes, as well as the questions on all latent constructs, were presented in randomized order. The latent constructs of the present study are: trust, distrust, privacy invasion, information disclosure, treatment adherence, satisfaction, as well as the individual difference variables: personality, disposition to trust humans, disposition to trust machines, technology attitude, and AI phobia. On the first page of the online questionnaire we provided a definition for the term "chatbot" (which was used in several item formulations) to secure a common understanding among the study participants and we also explained the study context.[5]

Data were collected in 2022 by a market research company.[6] Respondi, originally a German company, was recently acquired by the French market research company Bilendi,[7] but both brands continue to exist. Respondi provides services such as online surveys and access to diverse panels to support both academic research and market research. The company is ISO quality certified (ISO 20252:2019). The target population of the online experiment were individuals from the German-speaking area (Austria, Germany, Switzerland) with a minimum age of 18 years. All statistical analyses presented in the current paper are based on a total sample of $N = 1{,}183$ persons.[8] The median participation time in our online experiment was 29 min. Table 1 provides an overview of the main sample characteristics.

2.2 Measures

All latent constructs were measured based on psychometrically validated instruments which also showed high reliability in the present sample (Cronbach's Alpha > 0.7). Specifically, we used the following instruments with some wording adjustments to exactly fit the present research context: trust (Söllner et al., 2012) (5 items, sample item: "I have a good feeling when relying on the diagnosis decision."), distrust (Söllner et al., 2012) (5 items, sample item: "I have a bad feeling when relying on the treatment decision."), privacy invasion (Fischer et al., 2021) (4 items, sample item: "I fear that my interaction with the [physician/physician who is supported by the chatbot/chatbot] is less confidential than I would like it to be."), information disclosure (Bansal et al., 2010) (3 items, sample item: "I have a strong intention to disclose health information to the [physician/physician who is supported by the chatbot/chatbot]."), treatment adherence (DiMatteo et al., 1992) (4 items, sample item: "I will follow the [physician's/chatbot supported physician's/chatbot's] suggestions exactly."), satisfaction (Probst et al., 1997) (4 items, sample item: "In general, I would be satisfied with this type of diagnosis and treatment development method."), personality (Gosling et al., 2003) (big-five model, two items for each factor = 10 items, sample item for neuroticism: "I easily get nervous and insecure."), disposition to trust humans (Gefen, 2000) (4 items, sample item: "I generally have faith in humanity."), disposition to trust machines (Gefen, 2000) (4 items, sample item, "I tend to count upon machines."), technology attitude (Nam, 2019) (2 items, sample item, "Thinking about the possibility that computers and robots could do most of the work currently done by humans, how enthusiastic are you, if at all, about this possibility for society as a whole? not at all enthusiastic / not too enthusiastic / somewhat enthusiastic / very enthusiastic."), and AI phobia (Khasawneh, 2018) [7 items, sample item: "I am afraid of new technologies because one day they will make us (humans) obsolete."]. Appendix B summarizes the measurement instruments, including all items and Cronbach's Alphas.[9] Appendix C shows the correlation table with the six main factors investigated.

2.3 Data analyses

Based on SPSS 27 repeated measures analysis of variance (RM-ANOVA) was conducted to determine the influence of the two within-subject factors, patient's interaction partner (human doctor, human doctor with AI system, AI system only) and medical discipline (cardiology, orthopedics, dermatology, psychiatry) on trust, distrust, privacy invasion, information disclosure, treatment adherence, and satisfaction. For each dependent variable a separate RM-ANOVA was calculated. The description of the mean values and standard errors of the mean (SEM) values are visualized graphically in Figures 1–6 for each dependent variable and summarized in Appendix D. Tests of the within-subject effects and interaction effects are reported. To address violations of the sphericity assumption, the Greenhouse–Geisser correction to adjust the degrees of freedom was applied if the result of the Mauchly test reaches significance. Bonferroni tests were used as *post hoc* tests to correct for multiple comparisons. In addition to these RM-ANOVA models we calculated additional analyses of covariance (ANCOVAs) controlling for possible confounding effects (see Appendix E). Moreover, as outlined in Appendix F, based on LISREL v. 8.80 structural equation modeling (SEM) was applied. SEM was used to model and test that perceptions of trust, distrust, and privacy invasion predict information disclosure, treatment adherence, and satisfaction. The statistical foundations of our approach can be found in Jöreskog (1970); further details and statistical results are presented in Appendix F.

4 Regarding health status we measured, based on "yes" and "no" answers, the occurrence of illnesses in different domains such as orthopedics, dermatology, or internal medicine. Moreover, we measured whether a participant took medication on a regular basis.

5 Specifically, we explained that communication with a chatbot alone means to communicate with an automatic computer system (i.e., a machine).

6 https://www.respondi.com/

7 https://www.bilendi.de/

8 The data set was complemented by a small sample ($n = 97$) recruited by the authors via online advertisement. These 97 individuals are part of the original sample of $N = 1{,}188$. However, five persons were excluded because they did not specify some demographic data. Moreover, on the last page of the online questionnaire we asked the following question: "Did you answer the questionnaire carefully and conscientiously?" All 1,183 persons confirmed this question with "yes."

9 Note that calculation of reliability coefficients was not possible for personality and technology attitude because each scale only includes two items. Hence, we report intercorrelations for the items of each scale (see Appendix B).

 10.3389/fpsyg.2024.1422177

TABLE 1 Overview of main sample characteristics ($N = 1{,}183$).

Country (current residence)	Austria: 386 (33%), Germany: 441 (37%), Switzerland: 351 (30%), other: 1 (<1%)
Sex	Female: 648 (55%), male: 535 (45%)
Age (years)	18–19: 244 (21%), 30–39: 220 (19%), 40–49: 213 (19%), 50–59: 263 (23%), >60: 203 (18%)[1]
Educational attainment	No degree (6, <1%), compulsory education (47, 4%), apprenticeship / professional training (497, 42%), high school diploma (309, 26%), university degree (324, 28%)
Self-reported Health status	Generally chronically ill [yes: 711 (60%), no: 472 (40%)], mentally ill [yes: 135 (11%), no: 1,048 (89%)], orthopedically ill [yes: 135 (11%), no: 1,048 (89%)], dermatologically ill [yes: 70 (6%), no: 1,113 (94%)], illness in the domain of internal medicine [yes: 161 (14%), no: 1,022 (86%)], other illness [yes: 134 (11%), no: 1049 (89%)] (multiple answers possible)
Regular intake of medication[2]	Yes: 469 (40%), no: 713 (60%)

[1]Due to missing, incorrect or ambiguous information on age in the data set, only a sample of $N = 1{,}143$ was used for the description of the age groups as well as for analyses that include age.
[2]This also includes contraceptive.

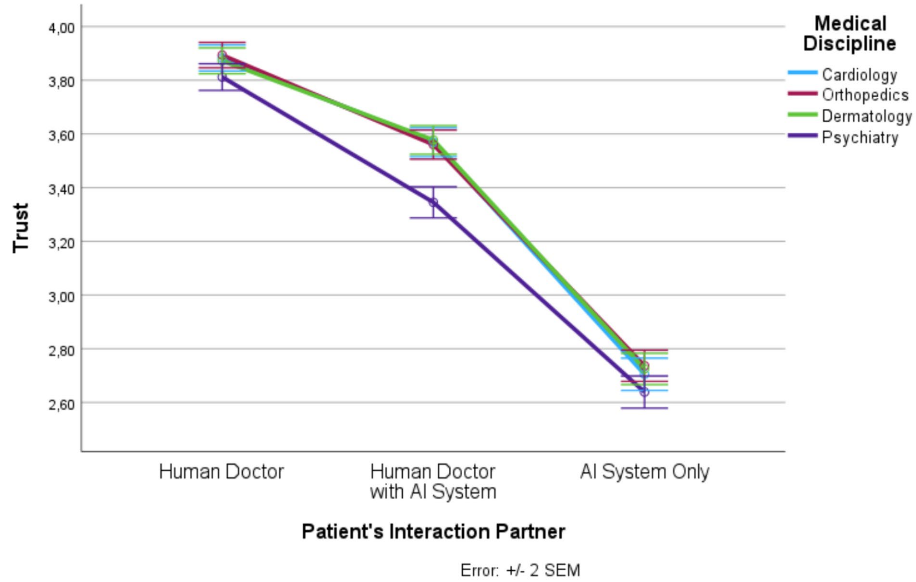


FIGURE 1
Patient's interaction partner and trust.

3 Results

3.1 Analysis of variance (RM-ANOVA)

In the following, we outline the RM-ANOVA results. The presentation is organized along the six factors, first showing the results for the main effect of the interaction partner, then the effects of the medical discipline, and finally we report the interaction effects between the two factors. Moreover, Figures 1–6 and Appendix D show the mean and SEM values for each factor.

3.1.1 Trust

Results show a strong main effect of the patient's interaction partner on trust (F (1.55, 1835.06) = 943.63, $p \leq 0.0001$, $\eta_p^2 = 0.444$; see Figure 1). In total, more than 44% of the variance in trust could be explained by the patient's interaction partner. We observed highest trust in the "human doctor" condition, lowest trust in the "AI system only" condition, and intermediate trust in the "human doctor with AI system" condition (F (1, 1,182) = 1179.57, $p \leq 0.0001$, $\eta^2 = 0.499$). All three levels differed significantly from each other.

Medical discipline explained about 5% of the variance in trust (F (2.77, 3277.31) = 64.64, $p \leq 0.0001$, $\eta_p^2 = 0.052$). Psychiatry had significantly lower trust scores than the other three disciplines. In addition, there was an interaction effect (F (5.76, 6812.50) = 11.10, $p \leq 0.0001$, $\eta_p^2 = 0.009$). For all levels of the patient's interaction partner, the trust scores were significantly lower in psychiatry than in the other three medical disciplines but this effect was strongest in the "human doctor with AI system" condition.

3.1.2 Distrust

Results show a strong main effect of the patient's interaction partner on distrust (F (1.55, 1830.39) = 799.16, $p \leq 0.0001$, $\eta_p^2 = 0.403$; see Figure 2). In total, 40% of the variance in distrust could be explained by the patient's interaction partner. We observed lowest distrust in the "human doctor" condition, highest distrust in the "AI system only" condition, and intermediate distrust in the "human

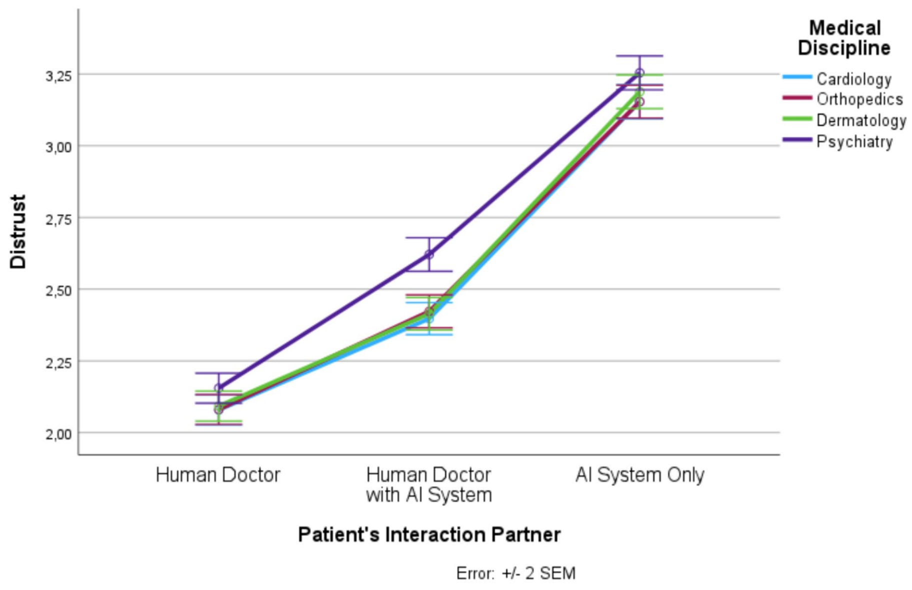


FIGURE 2
Patient's interaction partner and distrust.

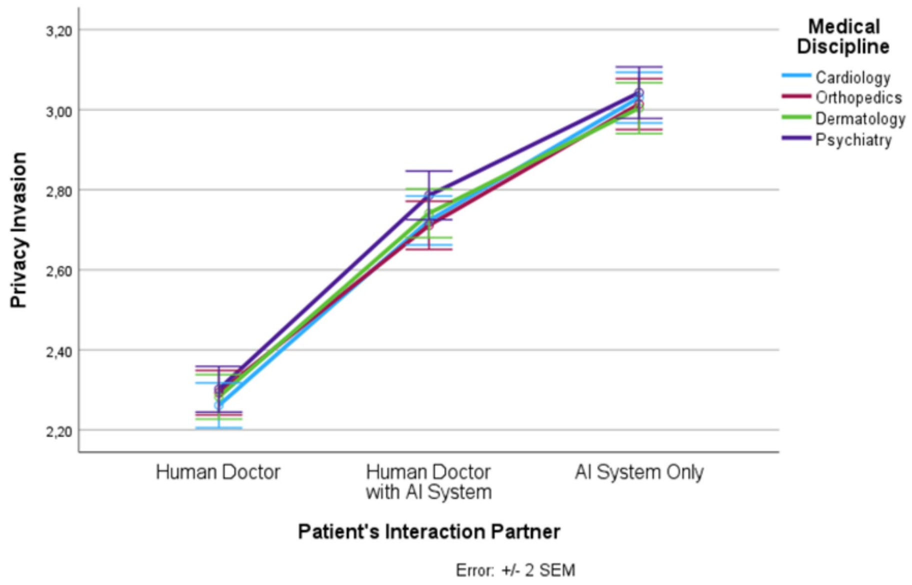


FIGURE 3
Patient's interaction partner and privacy invasion.

doctor with AI system" condition (F (1, 1,182) = 1012.41, $p \leq 0.0001$, $\eta_p^2 = 0.461$). All three levels differed significantly from each other.

Medical discipline explained about 4% of the variance in distrust (F (2.84, 3353.71) = 53.94, $p \leq 0.0001$, $\eta_p^2 = 0.044$). Psychiatry had significantly higher distrust scores than the other three disciplines. In addition, there was an interaction effect (F (5.77, 6820.03) = 7.65, $p \leq 0.0001$, $\eta_p^2 = 0.006$). For all levels of the patient's interaction partner, the distrust scores were significantly higher in psychiatry than in the other three medical disciplines but this effect was strongest in the "human doctor with AI system" condition.

3.1.3 Privacy invasion

Results show a strong main effect of the patient's interaction partner on privacy invasion (F (1.41, 1660.44) = 483.05, $p \leq 0.0001$, $\eta_p^2 = 0.290$; see Figure 3). In total, 29% of the variance in privacy invasion could be explained by the patient's interaction partner. We observed lowest privacy invasion in the "human doctor" condition, highest privacy invasion in the "AI system only" condition, and intermediate privacy invasion in the "human doctor with AI system" condition (F (1, 1,182) = 599.02, $p \leq 0.0001$, $\eta_p^2 = 0.336$). All three levels differed significantly from each other.

Medical discipline explained 0.5% of the variance in privacy invasion (F (3, 3,546) = 6.32, $p \leq 0.001$, $\eta_p^2 = 0.005$). Psychiatry had significantly higher privacy invasion scores than the other three disciplines. In addition, there was an interaction effect (F (5.93, 7010.83) = 2.12, $p = 0.049$, $\eta_p^2 = 0.002$). However, this interaction effect was only marginal. Privacy invasion was significantly higher for the medical discipline psychiatry compared with cardiology and orthopedics (but differed not significantly from dermatology), but this effect only occurred in the "human doctor with AI system" condition.

3.1.4 Information disclosure

Results show a strong main effect of the patient's interaction partner on information disclosure (F (1.46, 1725.46) = 590.69, $p \leq 0.0001$, $\eta_p^2 = 0.333$; see Figure 4). In total, more than 33% of the variance in information disclosure could be explained by the patient's interaction partner. We observed highest information disclosure in the "human doctor" condition, lowest information disclosure in the "AI system only" condition, and intermediate information disclosure in the "human doctor with AI system" condition (F (1, 1,182) = 726.64, $p \leq 0.0001$, $\eta_p^2 = 0.381$). All three levels differed significantly from each other.

Medical discipline explained about 3% of the variance in information disclosure (F (2.88, 3398.21) = 34.55, $p \leq 0.0001$, $\eta_p^2 = 0.028$). Psychiatry had significantly lower information disclosure scores than the other three disciplines. In addition, there was an interaction effect (F (5.85, 6909.81) = 4.16, $p \leq 0.001$, $\eta_p^2 = 0.004$). For all levels of the patient's interaction partner, the information disclosure scores were significantly lower in psychiatry than in the other three medical disciplines (except for the pairwise comparison between dermatology and psychiatry in the "human doctor" condition) but this effect was strongest in the "human doctor with AI system" condition.

3.1.5 Treatment adherence

Results show a strong main effect of the patient's interaction partner on treatment adherence (F (1.59, 1873.14) = 590.68, $p \leq 0.0001$, $\eta_p^2 = 0.333$; see Figure 5). In total, more than 33% of the variance in treatment adherence could be explained by the patient's interaction partner. We observed highest treatment adherence in the "human doctor" condition, lowest treatment adherence in the "AI system only" condition, and intermediate treatment adherence in the "human doctor with AI system" condition (F (1, 1,182) = 764.02, $p \leq 0.0001$, $\eta_p^2 = 0.393$). All three levels differed significantly from each other.

Medical discipline explained about 3% of the variance in treatment adherence (F (2.89, 3421.20) = 37.82, $p \leq 0.0001$, $\eta_p^2 = 0.031$). Psychiatry had significantly lower treatment adherence scores than the other three disciplines. In addition, there was an interaction effect (F (5.95, 7026.41) = 4.18, $p \leq 0.001$, $\eta_p^2 = 0.004$). For all levels of the patient's interaction partner, the treatment adherence scores were significantly lower in psychiatry than in the other three medical disciplines (except for the pairwise comparison between cardiology and psychiatry in the "AI system only" condition) but this effect was strongest in the "human doctor with AI system" condition.

3.1.6 Satisfaction

Results show a strong main effect of the patient's interaction partner on satisfaction (F (1.56, 1844.44) = 865.17, $p \leq 0.0001$, $\eta_p^2 = 0.423$; see Figure 6). In total, more than 42% of the variance in satisfaction could be explained by the patient's interaction partner. We observed highest satisfaction in the "human doctor" condition, lowest satisfaction in the "AI system only" condition, and intermediate satisfaction in the "human doctor with AI system" condition (F (1, 1,182) = 1085.348, $p \leq 0.0001$, $\eta_p^2 = 0.479$). All three levels differed significantly from each other.

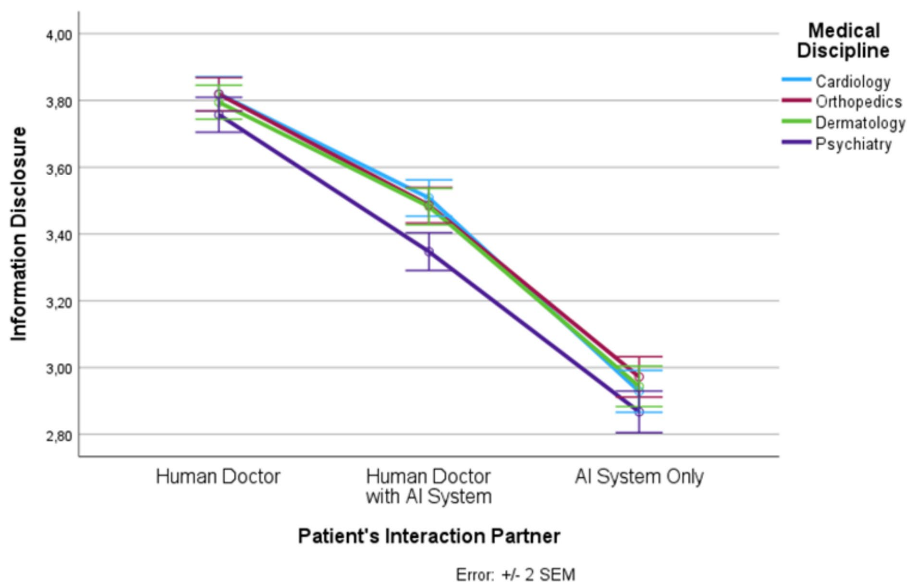


FIGURE 4
Patient's interaction partner and information disclosure.

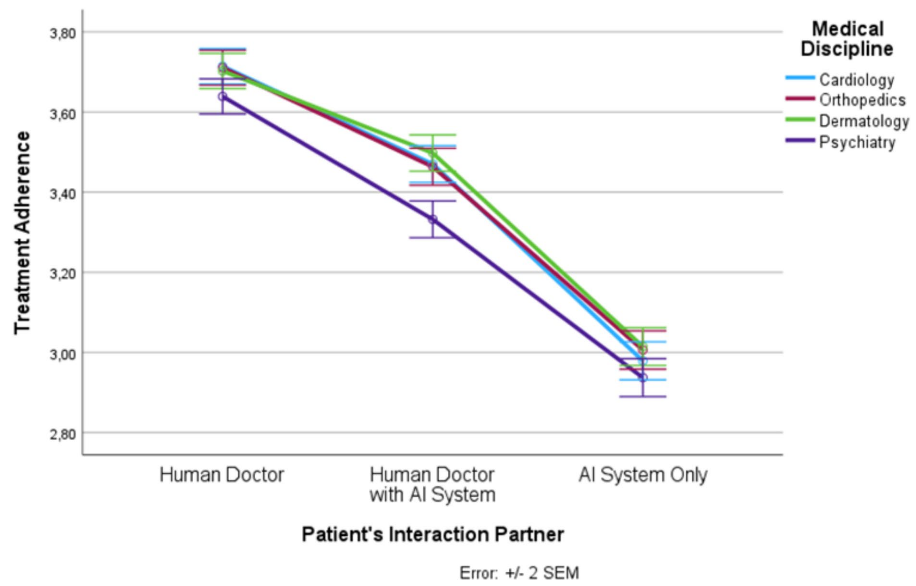


FIGURE 5
Patient's interaction partner and treatment adherence.

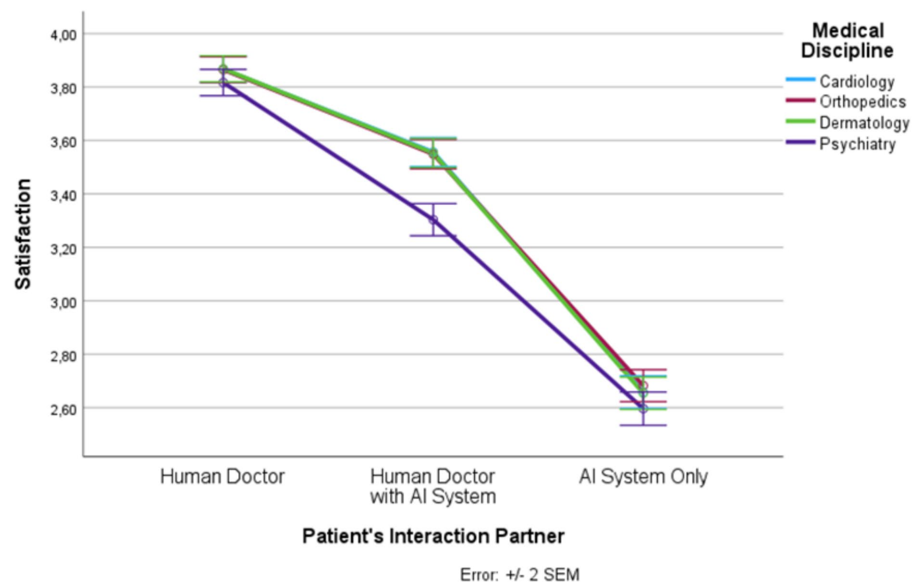


FIGURE 6
Patient's interaction partner and satisfaction.

Medical discipline explained 4.9% of the variance in satisfaction (F (2.76, 3263.35) = 60.85, $p \leq 0.0001$, $\eta_p^2 = 0.049$). Psychiatry had significantly lower satisfaction scores than the other three disciplines. In addition, there was an interaction effect (F (5.73, 6777.95) = 17.57, $p \leq 0.0001$, $\eta_p^2 = 0.015$). For all levels of the patient's interaction partner, the satisfaction scores were significantly lower in psychiatry than in the other three medical disciplines but this effect was strongest in the "human doctor with AI system" condition.

3.2 Structural equation model with latent variables

Evidence suggests that trust, along with distrust and privacy invasion, may serve as predictors of information disclosure, treatment adherence, and satisfaction [see, for example, the evidence reported in a meta-analysis by Birkhäuer et al. (2017)]. Given the theoretical background of the present study, we investigated the mentioned prediction as a function of the patient's interaction partner and

 10.3389/fpsyg.2024.1422177

TABLE 2 Total effects of predictors on dependent variables.

			Predictors		
			Trust	Distrust	Privacy invasion
Dependent variables	Information disclosure	β	1.03	0.36	−0.31
		SER	0.11	0.11	0.06
		t-value	9.23	3.36	−4.77
	Treatment adherence	β	0.45	−0.56	−0.04
		SER	0.08	0.12	0.06
		t-value	5.49	−4.72	−0.63
	Satisfaction	β	1.10	—	—
		SER	0.08	—	—
		t-value	13.96	—	—

The "—" sign in the table indicates that neither distrust nor privacy invasion has a statistically significant influence on satisfaction. Standardized regression coefficients: β. Standard errors of regression: SER.

medical discipline. To do this, we built a structural equation model (SEM). We calculated a mixed model, i.e., we combined measurement models of the dependent variables and predictors with the assessment of the relationship between their latent factors. Table 2 summarizes the standardized regression coefficients (β), standard errors of regression (SER), and the t-values.

Moreover, for R^2 we found the following results: information disclosure 0.99, treatment adherence 0.95, and satisfaction 0.24. What follows is that trust, distrust, and privacy invasion *together* explain significant variance in the three dependent variables; in case of information disclosure 99% and in case of treatment adherence 95%, a remarkable result, and in case of satisfaction they still explain approximately a quarter of the total variance (24%).

In addition to Table 2, we present further statistical results in Appendix F based on SEM and LISREL modeling.

4 Discussion

In the current study, we first investigated the question whether patients prefer a human doctor, a human doctor with an AI system, or an AI system alone. Thus, our study contributes to a better understanding of the acceptance of possible new forms of patient-doctor interaction and the consequences of a possible future shift from classical human-human interaction to interaction scenarios in which AI systems play a role, either as decision support for the doctor or as a completely autonomous chatbot (i.e., machine). We found that people prefer a human doctor, followed by a human doctor with an AI system, and an AI system alone came in last place. Specifically, trust, information disclosure, treatment adherence, and satisfaction were significantly higher and distrust and perceived privacy invasion were significantly lower in the "human doctor" condition than in the "human doctor with AI system" and "AI system alone" conditions. We observed the same result pattern when comparing the "human doctor with AI system" and the "AI system alone" conditions.

The implications of these findings for medical practice are far-reaching. Our results strongly suggest that patients prefer to interact with a human doctor *without* AI support. Therefore, while the enormous potential of AI systems in medical contexts such as automatic disease diagnosis or surgeries by robots is undeniable (e.g., Hashimoto et al., 2018; Kelly et al., 2019), interaction with a doctor who uses an AI system during the conversation with the patient, and even more direct interaction with an AI system without a human doctor at all, likely entails negative consequences and reduces the potential therapy success. Moreover, our results suggest that the negative attitude toward AI-supported, or AI-based, interaction in patient-doctor communication processes could be mediated by psychological variables related to trust (see Appendix F).

Our results also suggest that patients, if eventually enforced to interact with "intelligent" machines in the future (e.g., due to cost pressure in the health care sector), will likely engage in behaviors which could endanger their health, such as reduced information disclosure in anamnesis or reduced treatment adherence. Thus, substitution of human doctors by AI systems could ultimately result in higher costs (e.g., due to later diagnosis of diseases or because medication cannot be effective due to lack of treatment adherence). Moreover, based on our results, it is also likely that patient satisfaction will suffer when they have to interact with AI systems directly, or with a doctor who is supported by an AI system. Altogether, our data suggest that patients prefer direct face-to-face communication in their interaction with doctors.

Recent evidence indicates that human interaction partners are preferred to AI systems for different reasons: the ability to feel emotions (Bigman and Gray, 2018), to avoid perceptions of uniqueness neglect (Longoni et al., 2019), perceived care and value similarity (Yokoi et al., 2021), and the ability to demonstrate empathy and to be benevolent (Wu et al., 2022). Consistent with these reasons, research on classical human-computer interaction situations (e.g., user interacts with a desktop PC), as well as research in non-medical AI contexts such as autonomous vehicle use, found that "humanizing technology" (e.g., by simply giving the machine a human name or by endowing a machine with complex behaviors such as demonstrating empathy or reciprocity) may significantly increase trust in the technology, which typically positively affects acceptance and adoption of the technology (e.g., Brave et al., 2005; Waytz et al., 2014). Considering these findings, future research on patient interaction with an AI system could experimentally manipulate the machine's characteristics to resemble typical human attributes. The prediction is that trust in the AI system would increase (and we also expect reduced distrust and privacy invasion perceptions). However, research in this context must not ignore ethical principles. In particular, if empirical research confirms the prediction that "humanizing technology" positively affects its acceptance and adoption, a broad discussion in medicine, as well as in society at large, should follow before designers and engineers create AI-based technologies that influence humans in ethically questionable ways in situations where a human doctor is replaced by a machine.

Furthermore, our study revealed that the scores for the six outcome variables differed significantly between psychiatry and the three other disciplines (cardiology, orthopedics, dermatology). Specifically, we found lower scores for trust, information disclosure, treatment adherence, and satisfaction, and higher scores for distrust and privacy invasion. Notably, we observed an interaction effect of patient's interaction partner by medical discipline. This interaction effect was caused by the fact that the six outcome variables differed strongly between psychiatry and the three other disciplines in the "human doctor with AI system" condition, while this effect was not that strong in the other conditions (human doctor, AI system alone). This result suggests that the human doctor alone condition is appreciated and the AI system alone is disliked almost equally across all medical disciplines. However, in the hybrid condition (human doctor with AI system) the nature of the disease or disorder makes a difference. The diagnosis and treatment of psychiatric disorders, which are often not discussed openly (because a psychiatric problem is eventually more stigmatized than a physical illness), are based more on the direct interaction between doctor and patient (i.e., human-human interaction) than the diagnosis and treatment of illnesses in other medical disciplines. Also, the pathological thought patterns underlying psychiatric disorders are likely to be less clearly identifiable by AI algorithms, and less amenable to AI algorithmic influence, than diseases more directly related to body structures (e.g., heart, spine, skin). This could explain why patients dislike interaction with a doctor who is supported by an AI system more in psychiatry than in the other investigated disciplines, and this difference does not exist when a patient interacts with a human doctor or with an AI system only.

These results are consistent with the mentioned reasons why people prefer human interaction partners to AI systems. In psychiatry, human characteristics such as the ability to feel emotions, perceived care, or the ability to demonstrate empathy and to be benevolent are comparatively more important than in other medical disciplines because exchange of communicative stimuli and social cues are *the* processes of utmost importance in the interaction between the patient and the doctor, as signified, for example, by doctor-patient interaction in psychotherapy. It seems that people believe that when a psychiatrist uses an AI system as decision support during the interaction process, the doctor's attention shifts from the patient to the system, possibly reducing patient perceptions such as perceived care, empathy, and benevolence. In essence, patients with psychiatric conditions may prefer to interact with a human doctor for various reasons. The present study suggests that trust and human connection play an important role. Concerns about confidentiality and privacy may also influence preferences, and our data suggest that patients fear data breaches in the doctor with AI system and in the AI system only conditions more than in the human doctor condition.

Altogether, our finding on the interaction effect of a patient's interaction partner by medical discipline suggests a "stuck-in-the-middle" issue in psychiatry. The practical implication of this issue is that if possible, psychiatrist-patient interaction should take place face-to-face. If this is not possible or necessary, computer-mediated interaction (e.g., via videoconferencing) (e.g., Backhaus et al., 2012) could be chosen. Importantly, recent evidence shows that interaction with a fully automated AI chatbot could be an effective approach in the diagnosis and treatment of psychiatric disorders (e.g., Lee et al., 2022). This finding, however, is not supported by the results of our study because, at least based on the six factors examined, people clearly prefer a human doctor to a non-human chatbot. Future research is necessary to establish the exact reasons for differences in research findings. As a starting point, we recommend recent papers that address the ethics of using AI technology, particularly machine learning, in psychiatric settings (Starke et al., 2021, 2023), as well as related papers on topics such as machine learning in predicting treatment outcomes in psychiatry (Chekroud et al., 2021).

The present study has limitations, several of which could also be addressed in future research. First, our participants did not respond to interaction with a real doctor, a real doctor with AI system, or a real AI system. As a result of our vignette-based research approach, participants could eventually respond differently in the real world, if compared to our hypothetical scenarios. However, Robert et al. (2009) report evidence showing that results of vignette studies and non-vignette studies are similar in decision-making contexts, and based on this evidence they argue that "when individuals are presented with vignettes, they respond as they do in a real scenario [… yet] vignettes can never mirror completely the reality and dynamism of people's lives" (p. 267). Despite this limitation, we emphasize that the vignette-based approach constitutes an established method in the study of decision-making phenomena in the medical context (Gould, 1996; e.g., Hughes, 1998). Moreover, we emphasize that in the present study we did not examine learning of trust. Therefore, based on existing research on the foundations of trust learning (e.g., King-Casas et al., 2005; Krueger and Meyer-Lindenberg, 2019), future research must not ignore the effects of experiences in the interaction between doctor and patient on trust perceptions and the other variables examined in the present study.

Second, the sample of the present study was recruited in German-speaking countries (Austria, Germany, Switzerland). Because evidence indicates that trust is related to culture (e.g., Thanetsunthorn and Wuthisatian, 2019), a fact which also holds true specifically for trust in AI systems (e.g., Chien et al., 2016), future studies should test the generalizability of our findings to identify possible cultural differences.

Third, we found that most control variables only had a marginal influence on the main effect (see Appendix E). In particular, we highlight that this result is unexpected especially with respect to personality (Oksanen et al., 2020). However, personality can be conceptualized and measured in different ways. We conceptualized personality based on the Big Five and used Gosling et al.'s (2003) seminal short-version measurement instrument (i.e., TIPI = Ten Item Personality Measure). In addition to this measurement of universal personality traits, we measured four *specific* personality traits (disposition to trust humans, disposition to trust machines, technology attitude, AI phobia). However, a recent review paper at the nexus of trust in AI systems and personality analyzed $N = 58$ empirical articles and identified five conceptualizations of universal personality traits (e.g., HEXACO) and 33 specific personality traits (Riedl, 2022). What follows is that future research should use other personality conceptualizations and measurement instruments to replicate our finding that personality, in general, hardly affected the main effects (see our RM-ANOVA findings). In corresponding research efforts, scholars should consider a recent research agenda paper on the role of personality in AI systems and robot use (Matthews et al., 2021).

Fourth, another limitation of our study that warrants discussion is the potential for participants to have inaccurate or unjustified perceptions of AI technologies, which may have influenced their responses. Participants' expectations of AI and its capabilities are also shaped by external factors such as media portrayals and personal biases, leading to unrealistic expectations or undue skepticism. In our vignettes, while we provided a brief description of chatbots, we did not offer a comprehensive explanation of AI systems in general or Clinical Decision Support Systems (CDSS) and their potential roles in assisting doctors (we deliberately avoided this in the present study; we did not want to add to the already long length with additional explanations and definitions). However, this omission might have contributed to participants' perceptions and preferences being based on incomplete or incorrect information about AI systems and CDSS. Future research should aim to address this by setting more realistic expectations through more detailed descriptions and educational interventions about AI technologies, thereby allowing for a more informed evaluation of their potential impact on the doctor-patient relationship.

Fifth, we cannot rule out some degree of participant inattention. The study was relatively long, and participants had to provide information on the six outcome variables 12 times during the experiment. Although the order of presentation of the vignettes and latent constructs was randomized, which mitigates this potential problem, potential inattention could still affect the results. Given this limitation, potential future replication studies should include attention controls distributed throughout the study. We emphasize, however, that there is no reason to believe that some degree of participant inattention would systematically affect the main findings of the present study, including the main effect of interaction partner on trust, distrust, perceived privacy invasion, information disclosure, treatment adherence, and satisfaction, as well as the finding that the situation in psychiatry differs from that in cardiology, orthopedics, and dermatology.

Sixth and finally, regarding our manipulation "doctor with AI system" we highlight that the present experiment solely focused on the patient's perspective. Thus, it was not the goal of our study to also investigate the doctor's trust in the AI system (along with possible perceptions of distrust and privacy invasion). However, it is obvious that such a research focus is also critical in future research because the doctor's trust along with corresponding beliefs, attitudes, and behaviors could influence trust, distrust, perceived privacy invasion, information disclosure, treatment adherence, and patient satisfaction. As a starting point for corresponding research, we recommend a paper by Asan et al. (2020) who deal with the physician perspective in the study of trust in AI systems, as well as a paper by Jussupow et al. (2021) who empirically examined physicians' thought processes while using an AI system.

The most important implication of the current study's results is that *the replacement of a human doctor by an AI system in the doctor-patient discussion should not be carried out without taking into account the possible consequences*. Our conclusion is in line with statements in the academic literature on an ongoing debate regarding use of autonomous AI systems in medicine. Gille et al. (2020), for example, recently wrote that "[h]owever effective, these systems offer few clues as to how they arrive at their conclusions, hence raise questions of transparency, accountability and responsibility" (p. 1). Similarly, Powell (2019) argued that "many medical decisions require value judgments and the doctor-patient relationship requires empathy and understanding to arrive at a shared decision, often handling large areas of uncertainty and balancing competing risks. Arguably, medicine requires wisdom more than intelligence, artificial or otherwise. Artificial intelligence therefore needs to supplement rather than replace medical professionals, and identifying the complementary positioning of artificial intelligence in medical consultation is a key challenge for the future" (p. 1). It will be rewarding to see what insights future research will reveal.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

RR: Conceptualization, Funding acquisition, Methodology, Project administration, Validation, Writing – original draft. SH: Data curation, Formal analysis, Visualization, Writing – original draft. MR: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Austrian Science Fund (FWF) as part of the project "Technostress in Organizations" (project number: P 30865).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1422177/full#supplementary-material

References

Asan, O., Bayrak, A. E., and Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: focus on clinicians. *J. Med. Internet Res.* 22:e15154. doi: 10.2196/15154

Backhaus, A., Agha, Z., Maglione, M. L., Repp, A., Ross, B., Zuest, D., et al. (2012). Videoconferencing psychotherapy: a systematic review. *Psychol. Serv.* 9, 111–131. doi: 10.1037/a0027924

Bansal, G., Zahedi, F. ". M.", and Gefen, D. (2010). The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decis. Support. Syst.* 49, 138–150. doi: 10.1016/j.dss.2010.01.010

Barber, B. (1983). The logic and limits of trust. New Brunswick: Rutgers University Press.

Bigman, Y. E., and Gray, K. (2018). People are averse to machines making moral decisions. *Cognition* 181, 21–34. doi: 10.1016/j.cognition.2018.08.003

Birkhäuer, J., Gaab, J., Kossowsky, J., Hasler, S., Krummenacher, P., Werner, C., et al. (2017). Trust in the health care professional and health outcome: a meta-analysis. *PLoS One* 12:e0170988. doi: 10.1371/journal.pone.0170988

Brave, S., Nass, C., and Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *Int. J. Human-Comp. Stud.* 62, 161–178. doi: 10.1016/j.ijhcs.2004.11.002

Bruce, J. M., Hancock, L. M., Arnett, P., and Lynch, S. (2010). Treatment adherence in multiple sclerosis: association with emotional status, personality, and cognition. *J. Behav. Med.* 33, 219–227. doi: 10.1007/s10865-010-9247-y

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., et al. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* 20, 154–170. doi: 10.1002/wps.20882

Chien, S.-Y., Sycara, K., Liu, J.-S., and Kumru, A. (2016). Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. *Proc. Hum. Fac. Ergono. Soc. Annual Meet.* 60, 841–845. doi: 10.1177/1541931213601192

Dalton-Brown, S. (2020). The ethics of medical AI and the physician-patient relationship. *Camb. Q. Healthc. Ethics* 29, 115–121. doi: 10.1017/S0963180119000847

DiMatteo, M. R., Hays, R. D., and Sherbourne, C. D. (1992). Adherence to cancer regimens: implications for treating the older patient. *Oncology* 6, 50–57

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Q.* 34, 373–396. doi: 10.2307/20721433

Elgood, C. (2010). A medical history of Persia and the eastern caliphate: From the earliest times until the year a.D. 1932. Cambridge: University Press.

Fischer, T., Reuter, M., and Riedl, R. (2021). The digital stressors scale: development and validation of a new survey instrument to measure digital stress perceptions in the workplace context. *Front. Psychol.* 12:607598. doi: 10.3389/fpsyg.2021.607598

Gefen, D. (2000). E-commerce: the role of familiarity and trust. *Omega* 28, 725–737. doi: 10.1016/S0305-0483(00)00021-9

Gille, F., Jobin, A., and Ienca, M. (2020). What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intell. Based Med.* 1-2:100001. doi: 10.1016/j.ibmed.2020.100001

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the big-five personality domains. *J. Res. Pers.* 37, 504–528. doi: 10.1016/S0092-6566(03)00046-1

Gould, D. (1996). Using vignettes to collect data for nursing research studies: how valid are the findings? *J. Clin. Nurs.* 5, 207–212. doi: 10.1111/j.1365-2702.1996.tb00253.x

Hashimoto, D. A., Rosman, G., Rus, D., and Meireles, O. R. (2018). Artificial intelligence in surgery: promises and perils. *Ann. Surg.* 268, 70–76. doi: 10.1097/SLA.0000000000002693

Hollander, J. E., and Carr, B. G. (2020). Virtually perfect? Telemedicine for Covid-19. *N. Engl. J. Med.* 382, 1679–1681. doi: 10.1056/NEJMp2003539

Hughes, R. (1998). Considering the vignette technique and its application to a study of drug injecting and HIV risk and safer behaviour. *Sociol. Health Illn.* 20, 381–400. doi: 10.1111/1467-9566.00107

Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. *ETS Res. Bull. Series* 1970, i–41. doi: 10.1002/j.2333-8504.1970.tb00783.x

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Inf. Syst. Res.* 32, 713–735. doi: 10.1287/isre.2020.0980

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17:195. doi: 10.1186/s12916-019-1426-2

Khasawneh, O. Y. (2018). Technophobia: examining its hidden factors and defining it. *Technol. Soc.* 54, 93–100. doi: 10.1016/j.techsoc.2018.03.008

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83. doi: 10.1126/science.1108062

Krueger, F., and Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends Neurosci.* 42, 92–101. doi: 10.1016/j.tins.2018.10.004

Lee, J., Lee, D., and Lee, J. (2022). Influence of Rapport and Social Presence with an AI Psychotherapy Chatbot on Users' Self-Disclosure. *International Journal of Human–Computer Interaction*, 40, 1620–1631. doi: 10.1080/10447318.2022.2146227

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *J. Consum. Res.* 46, 629–650. doi: 10.1093/jcr/ucz013

Lewicki, R. J., McAllister, D. J., and Bies, R. J. (1998). "Trust and Distrust: New Relationships and Realities," *Academy of Management Review* 23, 438–558.

Mann, D. M., Chen, J., Chunara, R., Testa, P. A., and Nov, O. (2020). COVID-19 transforms health care through telemedicine: evidence from the field. *J. Am. Med. Inform. Assoc.* 27, 1132–1135. doi: 10.1093/jamia/ocaa072

Matthews, G., Hancock, P. A., Lin, J., Panganiban, A. R., Reinerman-Jones, L. E., Szalma, J. L., et al. (2021). Evolution and revolution: personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Personal. Individ. Differ.* 169:109969. doi: 10.1016/j.paid.2020.109969

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.2307/258792

McGrew, R. (1985). Encyclopedia of medical history. New York: McGraw-Hill Book Company.

McKnight, D. H., and Choudhury, V. (2006). Distrust and Trust in B2C e-Commerce: Do They Differ?. in *Proceedings of the 8th International Conference on Electronic Commerce: The New ECommerce: Innovations for Conquering Current Barriers, Obstacles, and Limitations to Conducting Successful Business on the Internet*, Fredericton, New Brunswick, Canada. pp. 482–491.

Nagy, M., and Sisk, B. (2020). How will artificial intelligence affect patient-clinician relationships? *AMA J. Ethics* 22, E395–E400. doi: 10.1001/amajethics.2020.395

Nam, T. (2019). Citizen attitudes about job replacement by robotic automation. *Futures* 109, 39–49. doi: 10.1016/j.futures.2019.04.005

Oksanen, A., Savela, N., Latikka, R., and Koivula, A. (2020). Trust toward robots and artificial intelligence: an experimental approach to human–technology interactions online. *Front. Psychol.* 11:568256. doi: 10.3389/fpsyg.2020.568256

Pearson, S. D., and Raeke, L. H. (2000). Patients' trust in physicians: many theories, few measures, and little data. *J. Gen. Intern. Med.* 15, 509–513. doi: 10.1046/j.1525-1497.2000.11002.x

Powell, J. (2019). Trust me, I'm a chatbot: how artificial intelligence in health care fails the turing test. *J. Med. Internet Res.* 21:e16222. doi: 10.2196/16222

Probst, J. C., Greenhouse, D. L., and Selassie, A. W. (1997). Patient and physician satisfaction with an outpatient care visit. *J. Fam. Pract.* 45, 418–425

Rai, A., Constantinides, P., and Sarker, S. (2019). Next generation digital platforms: toward human-AI hybrids. *MIS Q.* 43, iii–ix.

Riedl, R. (2022). Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electron. Mark.* 32, 2021–2051. doi: 10.1007/s12525-022-00594-4

Riedl, R., Hubert, M., and Kenning, P. (2010). Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *MIS Q.* 34, 397–428. doi: 10.2307/20721434

Riedl, R., and Javor, A. (2012). The biology of trust: integrating evidence from genetics, endocrinology, and functional brain imaging. *J. Neurosci. Psychol. Econ.* 5, 63–91. doi: 10.1037/a0026318

Robert, L. P., Dennis, A. R., and Hung, Y.-T. C. (2009). Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *J. Manag. Inf. Syst.* 26, 241–279. doi: 10.2753/MIS0742-1222260210

Söllner, M., Hoffmann, A., Hoffmann, H., Wacker, A., and Leimeister, J. M. (2012). Understanding the formation of trust in IT artifacts. Proceedings of the 31st international conference on information systems.

Spikins, P., Needham, A., Tilley, L., and Hitchens, G. (2018). Calculated or caring? Neanderthal healthcare in social context. *World Archaeol.* 50, 384–403. doi: 10.1080/00438243.2018.1433060

Starke, G., De Clercq, E., Borgwardt, S., and Elger, B. S. (2021). Computing schizophrenia: ethical challenges for machine learning in psychiatry. *Psychol. Med.* 51, 2515–2521. doi: 10.1017/S0033291720001683

Starke, G., Schmidt, B., De Clercq, E., and Elger, B. S. (2023). Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. *AI Ethics* 3, 303–314. doi: 10.1007/s43681-022-00177-1

Stewart, M. A. (1995). Effective physician-patient communication and health outcomes: a review. *Can. Med. Assoc. J.* 152, 1423–1433

Sutter, M., and Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Gam. Econ. Behav.* 59, 364–382. doi: 10.1016/j.geb.2006.07.006

Thanetsunthorn, N., and Wuthisatian, R. (2019). Understanding trust across cultures: an empirical investigation. *Rev. Int. Bus. Strat.* 29, 286–314. doi: 10.1108/RIBS-12-2018-0103

10.3389/fpsyg.2024.1422177

Vereschak, O., Bailly, G., and Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proc.ACM Hum. Comp. Inter.* 5, 1–39. doi: 10.1145/3476068

Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* 52, 113–117. doi: 10.1016/j.jesp.2014.01.005

Wu, Q., Jin, Z., and Wang, P. (2022). The relationship between the physician-patient relationship, physician empathy, and patient trust. *J. Gen. Intern. Med.* 37, 1388–1393. doi: 10.1007/s11606-021-07008-9

Yokoi, R., Eguchi, Y., Fujita, T., and Nakayachi, K. (2021). Artificial intelligence is trusted less than a doctor in medical treatment decisions: influence of perceived care and value similarity. *Int. J. Hum. Comp. Inter.* 37, 981–990. doi: 10.1080/10447318.2020.1861763

Supplementary Material

Do Patients Prefer a Human Doctor, Artificial Intelligence, or a Blend, and Is This Preference Dependent on Medical Discipline? Empirical Evidence and Implications for Medical Practice

René Riedl, Svea Hogeterp, Martin Reuter

Frontiers in Psychology, section Health Psychology (2024)

This *Supplementary Material* includes:

- Appendix A
- Appendix B
- Appendix C
- Appendix D
- Appendix E
- Appendix F

All references cited in this Supplementary Material are listed in the references of the main paper.

Appendix A

This appendix shows the introductory text and the twelve vignettes as used in the online experiment. The introductory text and the vignettes are also available in the original German language on request.

<u>Introductory text</u>

What will medical care look like in the future?

Digital technologies such as artificial intelligence (AI) and chatbots (i.e., computer systems that communicate with people automatically) are on the rise and are already being used in many areas of life (e.g., online shopping). Digital technologies are also becoming increasingly important in medicine. In this study, we want to investigate what attitudes and opinions people have about possible doctor-patient relationships in the future, and we are particularly interested in different degrees of digitalization in the doctor-patient relationship. Will we communicate directly with a doctor, will the doctor rely on computer-based decisions based on AI, or will we perhaps communicate exclusively with chatbots in the future?

This research project is a cooperative study between the Department of Psychology (Differential & Biological Psychology) at the University of Bonn, Germany, and the Austrian universities University of Applied Sciences Upper Austria, Campus Steyr, and the University Linz.

The survey takes about 30 minutes. Your participation is voluntary and anonymous and can be canceled at any time without giving reasons and without negative consequences. You must be at least 18 years old to participate.

By clicking on "Continue" you agree to participate in the survey and start the questionnaire.

Vignettes

Before the vignettes began, the following introductory text was presented:

We would now like to invite you to take part in a thought experiment: Please imagine that you are in the situations presented on the following pages. Below you will see statements on which you can express your opinion. You will go through a total of twelve different scenarios.

All twelve vignettes were started with the following introduction, followed by the text as shown below. After each vignette, participants were asked to answer the questions on trust, distrust, privacy invasion, information disclosure, treatment adherence, and satisfaction. The vignettes, as well as the questions on all latent constructs, were presented in randomized order.

A new health center has recently opened in the region where you live. Specialists from different disciplines as well as psychotherapists work in this health center. You have come to the health center because you have been suffering from medical conditions of unknown origin for some time.

Vignette 1 – Human doctor (cardiology)

You are in the examination room with a cardiologist. You tell him or her about your symptoms and a conversation about your symptoms develops. You name the following symptoms: cardiac arrhythmia, which occurs even with slight physical exertion, and shortness of breath. He or she will examine you, arrange for a resting ECG and take your blood. On the basis of a joint discussion and the ECG and laboratory results, the doctor makes the diagnosis and discusses the treatment with you.

Vignette 2 – Human doctor (orthopedics)

You are in the examination room with an orthopaedist. You tell him or her about your symptoms and a conversation about your symptoms develops. You name the following symptoms: severe back pain, which occurs particularly when sitting and standing. In addition, you tell him or her that the symptoms are more pronounced in the morning than in the evening. He or she will examine you and order an an x-ray of your spine. On the basis of a joint discussion and the x-ray findings, the doctor makes the diagnosis and discusses the treatment with you.

Vignette 3 – Human doctor (dermatology)

You are in the examination room with a dermatologist. You tell him or her about your symptoms and a conversation about your symptoms develops. You name the following symptoms: severe itching on your arms and legs that has persisted for several weeks. You also tell him or her that the scratching of the itchy areas has already caused bleeding. He or she will examine the itchy areas on your body and take a blood sample. On the basis of a joint discussion and the laboratory findings, the doctor makes the diagnosis and discusses the treatment with you.

Vignette 4 – Human doctor (psychiatry)

You are in the examination room with a psychiatrist. You tell him or her about your symptoms and a conversation about your symptoms develops. You name the following symptoms: feeling extremely listless for several months, increasing experience of anxiety, which is also accompanied by physical symptoms, in particular a racing heart, outbreaks of sweat and enormous tension and nervousness. He or she will ask you about your personal and working life. On the basis of a joint discussion, the doctor makes the diagnosis and discusses the treatment with you.

Vignette 5 – Human doctor with AI system (cardiology)

You are in the examination room with a cardiologist. You tell him or her about your symptoms and a conversation about your symptoms develops. You name the following symptoms: cardiac arrhythmia, which occurs even with slight physical exertion, and shortness of breath. He or she will examine you, arrange for a resting ECG and take your blood. A conversation with him or her follows. He or she informs you that a computer system based on artificial intelligence (AI) has automatically evaluated your reported symptoms together with the ECG and blood data, made a diagnosis and developed a treatment suggestion. The diagnosis and treatment suggestion are checked by the doctor. The doctor then discusses the diagnosis and treatment with you.

Vignette 6 – Human doctor with AI system (orthopedics)

You are in the examination room with an orthopaedist. You tell him or her about your symptoms and a conversation about your symptoms develops. You name the following symptoms: severe back pain, *which occurs particularly when sitting and standing. In addition, you tell him or her that the symptoms are more pronounced in the morning than in the evening. He or she will examine you and order an x-ray of your spine. A conversation with him or her follows. He or she informs you that a computer system that works on the basis of artificial intelligence (AI) has automatically evaluated your reported symptoms together with the X-ray images, made a diagnosis and developed a treatment suggestion. The diagnosis and treatment suggestion are checked by the doctor. The doctor then discusses the diagnosis and treatment with you.*

Vignette 7 – Human doctor with AI system (dermatology)

You are in the examination room with a dermatologist. You tell him or her about your symptoms and a conversation about your symptoms develops. You name the following symptoms: severe itching on your arms and legs that has persisted for several weeks. You also tell him or her that the scratching of the itchy areas has already caused bleeding. He or she will examine the itchy areas on your body and take a blood sample. A conversation with him or her follows. He or she informs you that a computer system that works on the basis of artificial intelligence (AI) has automatically evaluated your reported symptoms together with the blood data, made a diagnosis and developed a treatment suggestion. The diagnosis and treatment suggestion are checked by the doctor. The doctor then discusses the diagnosis and treatment with you.

Vignette 8 – Human doctor with AI system (psychiatry)

You are in the examination room with a psychiatrist. You tell him or her about your symptoms and a conversation about your symptoms develops. You name the following symptoms: feeling extremely listless for several months, increasing experience of anxiety, which is also accompanied by physical symptoms, in particular a racing heart, outbreaks of sweat and enormous tension and nervousness. He or she will ask you about your personal and working life. A conversation with him or her follows. He or she informs you that a computer system that works on the basis of artificial intelligence (AI) has automatically evaluated your reported symptoms and circumstances, made a diagnosis and developed a treatment suggestion. The diagnosis and treatment suggestion are checked by the doctor. The doctor then discusses the diagnosis and treatment with you.

Vignette 9 – AI system alone (cardiology)

You are in the examination room. There is a computer system and a robot, both of which work on the basis of artificial intelligence (AI). You communicate with the computer (chatbot) via a dialog box on the screen. You enter your symptoms on the keyboard, specifically cardiac arrhythmia, which occurs even with slight physical exertion, and shortness of breath. The system asks you questions to which you answer. The robot also takes a blood sample and performs an ECG. After some time, you will be presented with a diagnosis and treatment on the computer screen, which will later be implemented by a medical team.

Vignette 10 – AI system alone (orthopedics)

You are in the examination room. There is a computer system and a robot, both of which work on the basis of artificial intelligence (AI). You communicate with the computer (chatbot) via a dialog box on the screen. You enter your symptoms on the keyboard, specifically severe back pain, which occurs particularly when sitting and standing. In addition, you communicate that the symptoms are more pronounced in the morning than in the evening. The system asks you questions to which you answer.

An x-ray is also initiated. After some time, you will be presented with a diagnosis and treatment on the computer screen, which will later be implemented by a medical team.

Vignette 11 – AI system alone (dermatology)

You are in the examination room. There is a computer system and a robot, both of which work on the basis of artificial intelligence (AI). You communicate with the computer (chatbot) via a dialog box on the screen. You enter your symptoms on the keyboard, specifically severe itching on your arms and legs that has persisted for several weeks. You also communicate that the scratching of the itchy areas has already caused bleeding. The system asks you questions, which you answer. The robot also takes a blood sample. After some time, you will be presented with a diagnosis and treatment on the computer screen, which will later be implemented by a medical team.

Vignette 12 – AI system alone (psychiatry)

You are in the examination room. There is a computer system and a robot, both of which work on the basis of artificial intelligence (AI). You communicate with the computer (chatbot) via a dialog box on the screen. You enter your symptoms on the keyboard, specifically that you have been feeling extremely listless for several months and are increasingly experiencing anxiety, which is also accompanied by physical symptoms, in particular a racing heart, outbreaks of sweat and enormous tension and nervousness. The system asks you questions about your circumstances in relation to your private and work life, to which you answer. After some time, you will be presented with a diagnosis and treatment on the computer screen, which will later be implemented by a medical team.

Appendix B

This appendix summarizes the measurement instruments of the latent constructs, including all items and Cronbach's Alphas (α).

Table 1 Appendix B. Measurement instruments and Cronbach's Alphas (**α)**

Latent construct and items	**α or correlation between items**
Trust (Soellner et al. 2012) • The diagnosis decision is trustworthy. • The treatment decision is trustworthy. • I have a good feeling when relying on the diagnosis decision. • I have a good feeling when relying on the treatment decision. • I trust the information presented by the [physician / physician who is supported by the chatbot / chatbot]	.97
Distrust (Soellner et al. 2012, reverse formulation of trust items) • The diagnosis decision is untrustworthy. • The treatment decision is untrustworthy. • I have a bad feeling when relying on the diagnosis decision. • I have a bad feeling when relying on the treatment decision. • I distrust the information presented by the [physician / physician who is supported by the chatbot / chatbot]	.97
Privacy invasion (Fischer et al. 2021) • I fear that my interaction with the [physician / physician who is supported by the chatbot / chatbot] is less confidential than I would like it to be. • I fear that the information that I exchange with the [physician / physician who is supported by the chatbot / chatbot] is not as protected as I would like it to be. • My personal health information is too easily accessible due to my interaction with the [physician / physician who is supported by the chatbot / chatbot]. • I fear that my personal data can easily be stolen as a consequence of my interaction with the [physician / physician who is supported by the chatbot / chatbot].	.98

Information disclosure (Bansal et al. 2010) • I would be very likely to disclose health information to the [physician / physician who is supported by the chatbot / chatbot]. • I have a strong intention to disclose health information to the [physician / physician who is supported by the chatbot / chatbot]. • I would definitely disclose health information to the [physician / physician who is supported by the chatbot / chatbot].	.97
Treatment adherence (DiMatteo et al. 1992) • I will have a hard time doing what the [physician / physician who is supported by the chatbot / chatbot] suggested that I should do. (reverse coded) • I will follow the [physician's / chatbot supported physician's / chatbot's] suggestions exactly. • I will be unable to do what will be necessary to follow the [physician's / chatbot supported physician's / chatbot's] treatment plans. (reverse coded) • I will find it easy to do the things the [physician / physician who is supported by the chatbot / chatbot] suggested that I should do.	.94
Satisfaction (Probst et al. 1997) • This type of diagnosis and treatment development method seems to pay attention as I describe my condition. • This type of diagnosis and treatment development method handles me with care during the examination. • This type of diagnosis and treatment development method makes me feel as if I could talk about any type of problem. • In general, I would be satisfied with this type of diagnosis and treatment development method.	.96
Personality (based on Gosling et al. 2003) Extraversion • I am rather reserved, cautious. (reverse coded) • I come out of myself, am sociable. Agreeableness • I easily give others trust, believe in the good in people. • I tend to criticize others. (reverse coded) Conscientiousness • I make myself comfortable, tend toward laziness. (reverse coded) • I complete tasks thoroughly.	Correlations between items of each scale: E: $r = .51$ $p < .001$ A: $r = .07$ $p = .012$ C: $r = .20$ $p < .001$

Neuroticism • I am relaxed, do not let stress unsettle me. (reverse coded) • I easily get nervous and insecure. Openness • I have little artistic interest. (reverse coded) • I have an active imagination.	N: r = .43 p < .001 O: r = .32 p < .001
Disposition to trust humans (Gefen 2000) • I generally trust other people. • I tend to count upon other people. • I generally have faith in humanity. • I generally trust other people unless they give me reason not to.	.81
Disposition to trust machines (Gefen 2000, adjusted formulation of disposition to trust humans items) • I generally trust machines. • I tend to count upon machines. • I generally have faith in the effective operation of machines. • I generally trust machines unless they give me reason not to.	.78
Technology attitude (Nam 2019). This construct was measured with two sub-dimensions: enthusiasm and worry. Enthusiasm Thinking about the possibility that computers and robots could do most of the work currently done by humans, how enthusiastic are you, if at all, about this possibility for society as a whole? • not at all enthusiastic • not too enthusiastic • somewhat enthusiastic • very enthusiastic	Correlation between items: r = -.29 p < .001

Worry Thinking about the possibility that computers and robots could do most of the work currently done by humans, how worried are you, if at all, about this possibility for society as a whole? • not at all worried • not too worried • somewhat worried • very worried	
AI phobia (Khasanwneh 2018) • I am fearful that someone is using technology to watch and listen to everything that I do. • I am terrified that technologies will change the way we live, communicate, love, and even judge others. • I am afraid of new technologies because one day they will make us (humans) obsolete. • I am fearful that new technologies will someday take over my job. • I am afraid of new technologies because they may interfere with my life emotionally, physically, and psychologically. • I am afraid to use some features in my cell phone. • I am terrified of being connected to the Internet, because someone might be tracking me.	.85

Note: All latent constructs except technology attitude (Nam 2019) were measured with a 5-point scale, where participants expressed their agreement with each item. The scale ranged from 1 “not at all correct” to 5 “completely correct”.

Appendix C

This appendix shows the correlation table with the six main factors investigated.

		Trust	Distrust	Privacy Invasion	Information Disclosure	Treatment Adherence	Satisfaction
Trust	r	1					
	p						
Distrust	r	-.698	1				
	p	.000					
Privacy Invasion	r	-.459	.707	1			
	p	.000	.000				
Information Disclosure	r	.829	-.588	-.535	1		
	p	.000	.000	.000			
Treatment Adherence	r	.779	-.802	-.620	.734	1	
	p	.000	.000	.000	.000		
Satisfaction	r	.936	-.637	-.432	.843	.748	1
	p	.000	.000	.000	.000	.000	

Note: All correlations (r) are significant at $p < 0.01$ (two-tailed).

Appendix D

This appendix presents the means and standard errors of the means (SEM) for each of the six main factors investigated.

Trust

Interaction partner	Medical discipline	Mean	SEM
Human doctor	Cardiology	3.885	.024
	Orthopaedics	3.894	.024
	Dermatology	3.874	.024
	Psychiatry	3.812	.025
Human doctor with AI	Cardiology	3.572	.027
	Orthopaedics	3.563	.027
	Dermatology	3.580	.027
	Psychiatry	3.348	.029
AI system only	Cardiology	2.708	.030
	Orthopaedics	2.739	.029
	Dermatology	2.727	.029
	Psychiatry	2.642	.030

Distrust

Interaction partner	Medical discipline	Mean	SEM
Human doctor	Cardiology	2.077	.026
	Orthopaedics	2.079	.026
	Dermatology	2.091	.026
	Psychiatry	2.154	.026
Human doctor with AI	Cardiology	2.394	.028
	Orthopaedics	2.420	.029
	Dermatology	2.411	.028
	Psychiatry	2.618	.029
AI system only	Cardiology	3.151	.030
	Orthopaedics	3.151	.029
	Dermatology	3.187	.029
	Psychiatry	3.252	.030

Privacy Invasion

Interaction partner	Medical discipline	Mean	SEM
Human doctor	Cardiology	2.261	.028
	Orthopaedics	2.292	.028
	Dermatology	2.282	.028
	Psychiatry	2.301	.029
Human doctor with AI	Cardiology	2.722	.030
	Orthopaedics	2.711	.030
	Dermatology	2.740	.030
	Psychiatry	2.786	.030
AI system only	Cardiology	3.031	.032
	Orthopaedics	3.014	.032
	Dermatology	3.005	.032
	Psychiatry	3.042	.032

Information Disclosure

Interaction partner	Medical discipline	Mean	SEM
Human doctor	Cardiology	3.820	.026
	Orthopaedics	3.818	.025
	Dermatology	3.797	.025
	Psychiatry	3.756	.026
Human doctor with AI	Cardiology	3.510	.027
	Orthopaedics	3.488	.027
	Dermatology	3.482	.027
	Psychiatry	3.348	.028
AI system only	Cardiology	2.931	.031
	Orthopaedics	2.975	.030
	Dermatology	2.947	.030
	Psychiatry	2.869	.031

Treatment Adherence

Interaction partner	Medical discipline	Mean	SEM
Human doctor	Cardiology	3.714	.022
	Orthopaedics	3.710	.022
	Dermatology	3.704	.022
	Psychiatry	3.637	.022
Human doctor with AI	Cardiology	3.472	.023
	Orthopaedics	3.464	.023
	Dermatology	3.499	.023
	Psychiatry	3.334	.023
AI system only	Cardiology	2.980	.024
	Orthopaedics	3.007	.024
	Dermatology	3.016	.024
	Psychiatry	2.939	.024

Satisfaction

Interaction partner	Medical discipline	Mean	SEM
Human doctor	Cardiology	3.868	.025
	Orthopaedics	3.866	.024
	Dermatology	3.870	.024
	Psychiatry	3.816	.024
Human doctor with AI	Cardiology	3.560	.027
	Orthopaedics	3.551	.027
	Dermatology	3.556	.027
	Psychiatry	3.306	.030
AI system only	Cardiology	2.660	.031
	Orthopaedics	2.686	.030
	Dermatology	2.657	.030
	Psychiatry	2.600	.031

Appendix E

In this appendix we present the results of our calculations regarding the control variables. The covariates were z-transformed before being entered in the ANCOVA models. For each control variable a separate ANCOVA model was computed. The list of the control variables is as follows: personality (extraversion, neuroticism, openness, agreeableness, and conscientiousness), disposition to trust humans, disposition to trust machines, AI phobia, and technology attitude (enthusiasm and worry). Further categorial control variables (chronic illness, regular medication intake, sex, education and age-group) were tested in separate additional ANOVA models.

With respect to the control variables which we entered as covariates into the ANCOVA models (personality with its dimensions extraversion, neuroticism, openness, agreeableness, and conscientiousness; disposition to trust humans; disposition to trust machines; AI phobia; and technology attitude with its dimensions enthusiasm and worry), we see in part substantial effects on the six major outcome variables (see Table 1 Appendix E, upper part). However, all covariate main or interaction effects are less strong than the effects of the patient's interaction partner. Moreover, these effects were mostly expected and hence are not surprising. For example, disposition to trust machines explains a substantial amount of variance in trust (η_p^2 = .167).

Additional ANOVA models including chronic illness, regular medication intake, sex, education, and age group as additional between-subject factors revealed occasional main or interaction effects; however, these effects were marginal in comparison to the effects of the patient's interaction partner and medical discipline (i.e., the two manipulated factors; see Table 1 Appendix E, lower part).

Table 1 Appendix E. Presentation of the strongest effects of control variables in terms of explained variance (partial eta^2)

	Trust	Distrust	Privacy Invasion	Information Disclosure	Treatment Adherence	Satisfaction
Covariates						
Extraversion	.010§	.009§	.002§	.010§	.010§	.013§
Neuroticism	.008§	.010§	.008§	.005§	.008§	.014§
Openness	.007$^{\$}$	.016§	.008$^{\$}$	.011§	.029§	.010$^{\$}$
Agreeableness	.023§	.019§	.010§	.017§	.014§	.022§
Conscientiousness	.015$^{\$}$	.026§	.012$^{\$}$	.019$^{\$}$	.028§	.019$^{\$}$
Disposition to Trust humans	.066§	.010§	.006§	.054§	.021§	.054§
Disposition to trust machines	.167§	.055§	.049§	.148§	.105§	.157§
AI phobia	.065§	.176§	.209§	.077§	.137§	.052§
Technology attitude: Enthusiasm	.081§	.039$^{\$}$	.030$^{\$}$	.079§	.049$^{\$}$	.074§
Technology attitude: Worry	.073§	.067§	.073§	.069§	.070§	.062§
Between-subject factors						
Chronic illness	.003*	.001$^{\$\#*}$	.001$^{\#}$	.001$^{\%\#}$	.003$^{\$\#}$	.001$^{\$\%\#*}$
Regular medication intake	.004$^{\%}$	.004$^{\$}$	.002$^{\$}$	.003$^{\$}$	.006$^{\$}$	.003$^{\$\%\#*}$
Sex	.009$^{\$}$	.009$^{\$}$	.007$^{\$}$	.007$^{\$}$	.013$^{\$}$	.015$^{\$}$
Education	.011*	.014*	.014*	.019*	.018*	.016$^{\$}$
Age group	.005$^{\$\%}$	.005$^{\$*}$	.007$^{\$}$	.008$^{\$}$	.007$^{\$}$	.007$^{\$}$

Note. All covariates were z-transformed before their inclusion in the model. § Main effect covariate, * main effect between-subject factor, $^{\$}$ interaction with patient's interaction partner, $^{\%}$ interaction with medical discipline, $^{\#}$ interaction with patient's interaction partner x medical discipline.

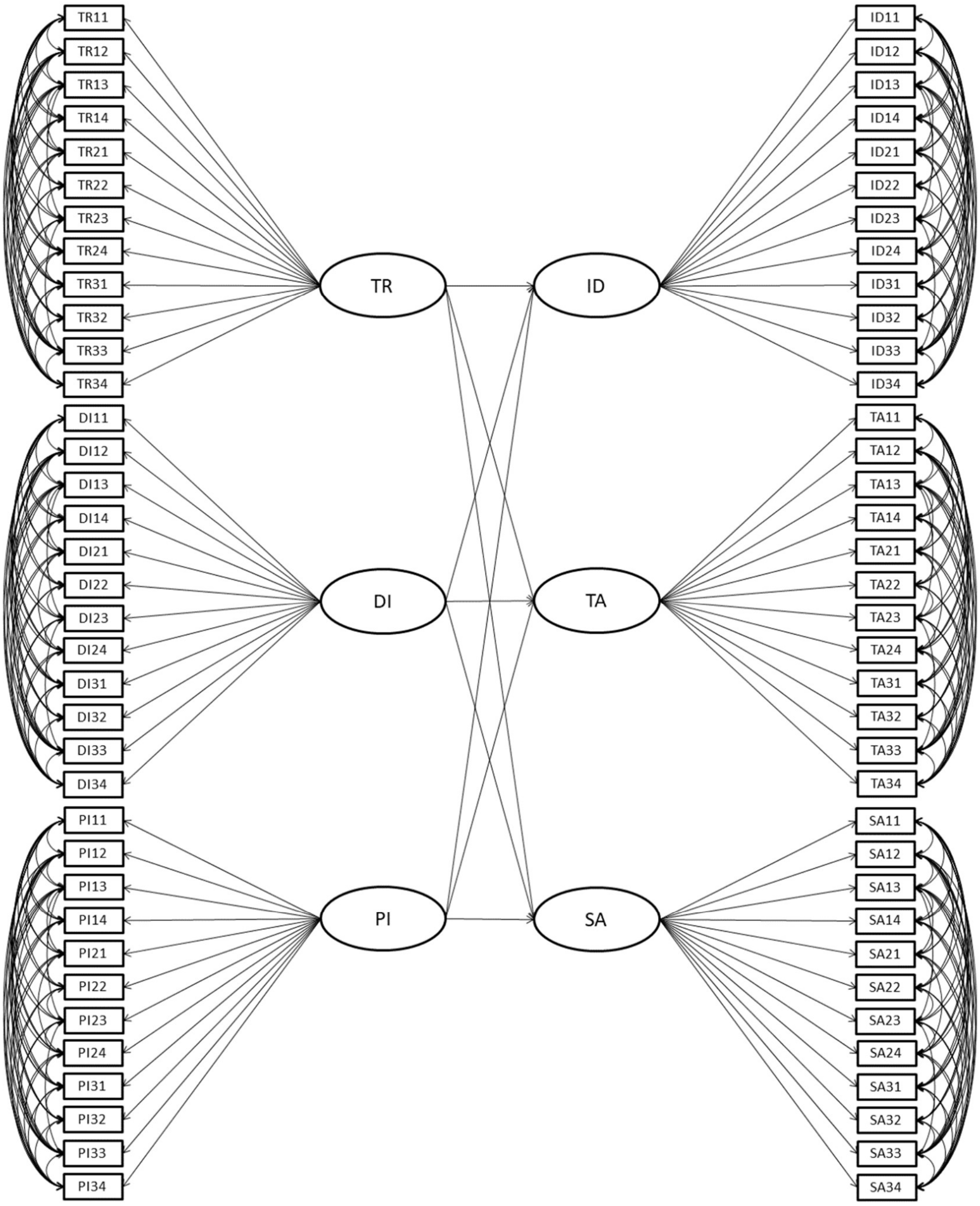
Appendix F

In this appendix we present our LISREL model. In this model all measurement items─lambda-X (LX) and lambda-Y (LY)─of trust (TR, 12 items), distrust (DI, 12 items), and privacy invasion (PI, 12 items) (formally expressed as TR/DI/PI_LXij with i = number of patient's interaction partner (i = 1, 2, 3) and j = number of medical discipline (j = 1, 2, 3, 4)), as well as all measurement items of information disclosure (ID, 3 items), treatment adherence (TA, 4 items), and satisfaction (SA, 4 items) (formally expressed as ID/TA/SA_LYij with i = number of patient's interaction partner (i = 1, 2, 3) and j = number of medical discipline (j = 1, 2, 3, 4)), represent the observed variables. Furthermore, the mediators (trust, distrust, privacy invasion) (Ksi formally expressed as $\xi 1$, $\xi 2$, $\xi 3$), as well as the dependent variables (information disclosure, treatment adherence, satisfaction) (Eta formally expressed as $\eta 1$, $\eta 2$, $\eta 3$), represent the latent variables. In our model, we specify: $\xi 1$ = TR, $\xi 2$ = DI, $\xi 3$ = PI; $\eta 1$ = ID, $\eta 2$ = TA, $\eta 3$ = SA; i1 = human doctor, i2 = human doctor with AI system, i3 = AI system only; j1 = cardiology, j2 = orthopedics, j3 = dermatology, j4 = psychiatry. The statistical foundations of our approach can be found in Jöreskog (1970). In the following Results section, we present a graphical representation of our LISREL model with the corresponding statistics.

Figure 1 Appendix F shows a conceptual representation of our LISREL model. The corresponding empirical path diagram is shown in Figure 2 Appendix F; the corresponding standardized path coefficients are summarized in Tables 1 and 2 Appendix F. Because both distrust and privacy invasion do not have a statistically significant influence on satisfaction (see Table 2 in the main manuscript), the corresponding paths (i.e., DI and PI to SA) are not shown. Regarding specification of the LISREL model, it can be seen that the six variables (trust, distrust, privacy invasion, information disclosure, treatment adherence, satisfaction) are conceptualized as reflective constructs.[1]

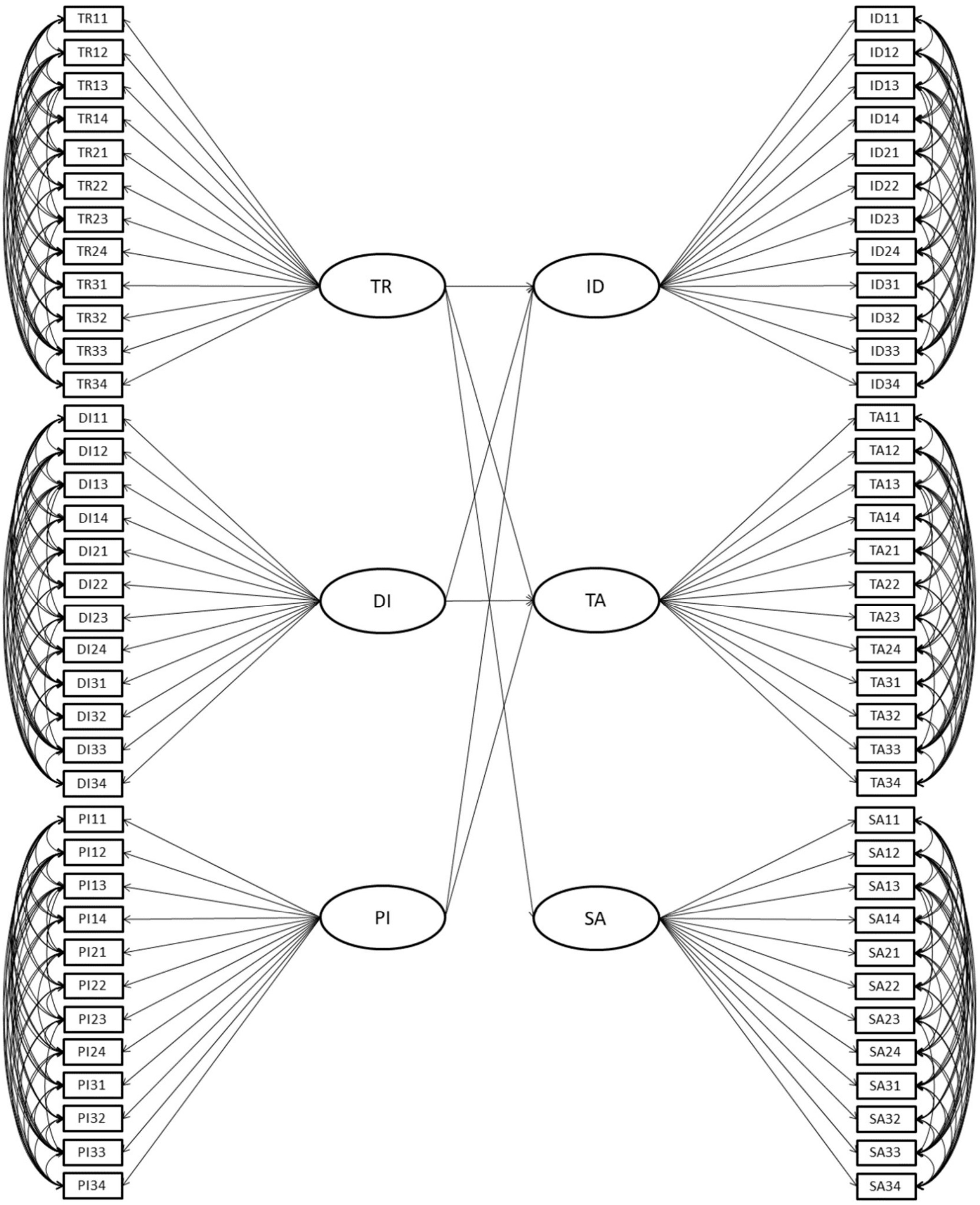
[1] Thus, the indicators (i.e., the items in the questionnaire) are caused by the latent variables (i.e., trust, distrust, privacy invasion, information disclosure, treatment adherence, satisfaction).

Figure 1 Appendix F. Conceptual Representation of the LISREL Model



Note: TR = Trust, DI = Distrust, PI = Privacy Invasion, ID = Information Disclosure, TA = Treatment Adherence, SA = Satisfaction.

Figure 2 Appendix F. Empirical Path Diagram



Note: TR = Trust, DI = Distrust, PI = Privacy Invasion, ID = Information Disclosure, TA = Treatment Adherence, SA = Satisfaction.

In the following, we indicate the standardized path coefficients of the LISREL model as shown in Figure 2 Appendix F.

Table 1 Appendix F. Values of lambda-X

	TR	DI	PI
TR11	0.41		
TR12	0.39		
TR13	0.41		
TR14	0.40		
TR21	0.61		
TR22	0.61		
TR23	0.61		
TR24	0.56		
TR31	0.35		
TR32	0.36		
TR33	0.34		
TR34	0.32		
DI11		0.40	
DI12		0.37	
DI13		0.42	
DI14		0.40	
DI21		0.59	
DI22		0.61	
DI23		0.60	
DI24		0.56	
DI31		0.33	
DI32		0.35	
DI33		0.34	
DI34		0.32	
PI11			0.43
PI12			0.42
PI13			0.42
PI14			0.43
PI21			0.65
PI22			0.66
PI23			0.66
PI24			0.65
PI31			0.55
PI32			0.54
PI33			0.54
PI34			0.53

Note: TR = Trust, DI = Distrust, PI = Privacy Invasion.

Table 2 Appendix F. Values of lambda-Y

	ID	TA	SA
ID11	0.35		
ID12	0.43		
ID13	0.45		
ID14	0.35		
ID21	0.64		
ID22	0.64		
ID23	0.66		
ID24	0.62		
ID31	0.43		
ID32	0.43		
ID33	0.44		
ID34	0.42		
TA11		0.25	
TA12		0.38	
TA13		0.37	
TA14		0.37	
TA21		0.58	
TA22		0.57	
TA23		0.57	
TA24		0.54	
TA31		0.38	
TA32		0.38	
TA33		0.39	
TA34		0.35	
SA11			0.37
SA12			0.36
SA13			0.37
SA14			0.37
SA21			0.57
SA22			0.56
SA23			0.56
SA24			0.49
SA31			0.29
SA32			0.29
SA33			0.27
SA34			0.26

Note: ID = Information Disclosure, TA = Treatment Adherence, SA = Satisfaction.