Essays in Behavioral Microeconomic Theory

Inauguraldissertation

zur Erlangung des Grades eines Doktors
der Wirtschaftswissenschaften

durch die

Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität

Bonn

vorgelegt von

Andreas Klümper

aus Dorsten

Bonn 2021

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Matthias Kräkel
Zweitreferent:	Prof. Dr. Francesc Dilmé
Tag der mündlichen Prüfung:	25. Juni 2021

Acknowledgements

Writing this dissertation would not have been possible without the incredible support of many people around me. As it is impossible to name them all in these paragraphs, I would like to let those who will not be named know that I am very grateful for their help.

Especially, I would like to thank my supervisor Professor Matthias Kräkel. With his patience, guidance, and feedback, he supported me in every step during the writing process. Whenever the difficulties of a dissertation seemed to be overwhelming, he was a true mentor by giving me the advice that I needed. For his help and support, I will always be thankful.

I am also grateful for the support of Professor Francesc Dilmé, who gave valuable comments on all of my projects and acted as a second referee on the thesis committee. Moreover, I would like to thank Professor Florian Ederer for his advice during my research visit to Yale University. His advice gave me the necessary additional motivation for the final push.

While writing my thesis, my friends, colleagues, and coauthors, Simon Dato, Andreas Grunewald, Sebastian Schaube, and Fabian Schmitz, helped me with their endless patience, friendliness, and detailed feedback. While I appreciated our professional and research-related discussions, coffee breaks and lunchtimes would not have been the same without our conversations about the recent football matches or the transfer market.

Besides my advisors and colleagues, I would like to thank the Bonn Graduate School of Economics and the BRIQ Institute on Behavior and Inequality, particularly Britta Altenburg, Armin Falk, Silke Kinzig, Benny Moldovanu, and Andrea Reykers,

for their efforts in providing an excellent research environment. Also, I would like to thank my fellow students at the Bonn Graduate School of Economics, from whom I have learned a lot and who made the extensive studying in the first years much more enjoyable.

While numerous people played a significant role in my dissertation's success, I would never have been able to start this journey without the endless support of my parents, Sabine and Manfred Klümper. Their excellent parenting allowed me to grow and gave me the self-confidence to begin this journey. When I was not sure about my way, they encouraged me to move on and supported me by all means possible. Together with my siblings, Markus and Claudia, my parents formed the foundation for this dissertation.

For me, my most important supporter was my wife, partner, and friend, Lisa. She supported me in every way one can think of. I will always be grateful, not only for her understanding but her encouragement to visit Yale University directly after our wedding. Lisa, the best decision I have ever made was to ask you to marry me, and it still makes me happy that you agreed. I owe you a lot for your endless support, patience, and unconditional love during the last more than eleven years. As long as we move on together, I am looking forward to all things that will cross our way. I genuinely love you and always will.

Contents

List of Figures

Introduction

The traditional economic view about human behavior, often referred to as the concept of *homo economicus*, relies upon the premise that individuals make rational choices. It assumes that they correctly process all the information available and have a deep understanding of their actions' consequences. Because the analysis of the homo economicus provides crucial insights about rational behavior in economic environments, economists do not disregard the included assumptions lightly.

However, psychologists and experimental economists have documented numerous departures from fully rational behavior. These observations have rendered the concept of the homo economicus as too optimistic regarding humans' cognitive abilities. Consequently, acknowledging that some economic actors do not behave fully rational, behavioral economists replaced the traditional concept. By combining psychology and economics, they have started to investigate the consequences of such behavior for individual decision-making and strategic interactions. Behavioral Economic Theory provides the toolbox for modeling departures from rational behavior. Therefore, it allows economists to make predictions about whether the existence of non-rational behavior changes the allocation of scarce resources, the distribution of economic welfare, and the impact of policy measures that aim to improve efficiency or redistribute welfare.

In the four chapters of this dissertation, I contribute to this line of research by analyzing the consequences of well-established departures from rationality and their remedies in strategic interactions. More precisely, I investigate credulity, reduced sensitivity to partitioned prices, and correlation neglect in models of democratic elections, imperfect competition, and contracting under hidden action. Moreover,

to facilitate the analysis of other departures from rationality and their impact on economic well-being, I define a solution concept for games between players with dual-selves and discuss its relations to existing equilibrium concepts in Behavioral Economic Theory.

The first chapter of this dissertation, Chapter 1, lays down how information manipulation during a democratic election, e.g., via spreading fake news, affects the quality of candidate selection and the polarization of political attitudes in the electorate. In the model, the electorate observes signals that political candidates potentially distort. Manipulation matters in equilibrium because a fraction of the electorate naively takes signals at face value. The results show that information manipulation increases the polarization of political attitudes. However, its effect on candidate selection is more subtle: information manipulation will be detrimental to candidate selection only if one of the candidates is substantially more capable of manipulating information. Otherwise, it will improve candidate selection. Finally, it is shown under what conditions policies that aim to equalize candidates' access to the media can help to attenuate detrimental effects of information manipulation in democratic elections. In the model, equalizing the access to the media can be interpreted as equalizing the cost to manipulate information between the candidates. If costs are relatively heterogeneous, decreasing their heterogeneity improves candidate selection and reduces the polarization of political attitudes. However, when costs are already relatively homogeneous, making costs even more homogeneous might impair candidate selection while the polarization of political attitudes still decreases. In the second case, whether leveling the playing field between the two candidates by equalizing their access to the media increases or decreases welfare depends on the weight of political polarization in the welfare function.

Shifting the focus from political institutions to markets for products, Chapter 2 discusses the impact of departures from rationality and their potential remedies in product markets. More precisely, it studies how partitioned pricing, i.e., the practice to split a good's price into a headline and an additional price, and its regulation affects consumer surplus and welfare. Because additional prices are less transparent than headline prices in this model, consumers react less sensitive to price changes

in the additional prices than to changes in the headline prices. Exploiting the consumers' naivety, firms use partitioned pricing to sell higher quantities and charge higher prices. Consequently, partitioned pricing benefits firms on the costs of consumers. The model allows us to analyze the impact of two measures of consumer protection that aim at reducing the exploitative nature of partitioned pricing: increasing the transparency of additional prices and regulating additional prices by determining a maximal value that additional prices may not exceed. Independent of the considered measure, the model shows that in markets with imperfect competition, a fundamental trade-off between consumer protection and efficiency arises. Full consumer protection, i.e., making prices fully transparent or prohibiting partitioned pricing, maximizes consumer surplus but leads to inefficiently low demand. Therefore, it is never welfare optimal in markets with imperfect competition. However, in markets with imperfect competition and unrestricted headline prices, any upper bound on the additional price increases welfare and consumer surplus compared to an unregulated additional price. In markets with non-negative headline prices, capping additional prices at the firms' costs is welfare neutral and increases consumer surplus. The chapter concludes by analyzing the interplay between the welfare-optimal degree of price transparency and regulation of additional prices.

Chapter 3 studies the effect of correlation neglect on a contracting problem between a principal and an agent when the principal wants to induce the agent to exert effort but cannot observe the agent's effort choice directly (hidden action). Therefore, it analyzes the welfare consequences if the agent neglects the correlation between performance measures in the linear Holmström-Milgrom model. It shows that the results on material welfare, the principal's expected profit, and the agent's expected utility are ambiguous. Whenever the true correlation and the perceived correlation between the performance measures are sufficiently low, the principal can impose a higher risk on the agent. Therefore, correlation neglect enhances material welfare and the principal's expected profit in this case. Conversely, when the perceived correlation between the performance measures is high, material welfare goes down, and the agent benefits from correlation neglect due to higher insurance.

While the first three chapters discuss the consequences of behavioral biases in specific applications, Chapter 4 considers a general class of games in which two selves with potentially different objectives govern each player's beliefs and actions. As human decision-making often seems to be determined by intrapersonal conflict resolution, the chapter conceptualizes the analysis of decisions governed by such dual-self processes in individual decision contexts and strategic interactions. For this purpose, it defines the solution concept Dual-Selves equilibrium and derives sufficient conditions for the existence of such equilibria. It shows that this result also helps to study the strategic interaction between players that have only one self but are not fully rational. In particular, the results extend several equilibrium existence results in Behavioral Game Theory and provide simple sufficient conditions for equilibrium existence in games between single-self players that are not fully rational.

Chapter 1

Welfare Effects of Information Manipulation in Political Elections*

Joint with Andreas Grunewald and Matthias Kräkel

1.1 Introduction

Representative democracy builds upon the premise that voters elect suitable political candidates to make decisions on their behalf. Voters' ability to screen politicians crucially depends on the quality of the available information. The recent spread of social media, however, came along with an increase of false information in political campaigns due to faked statistics, interferences of foreign countries, and false stories (Allcott and Gentzkow, 2017; Lazer, Baum, Benkler, Berinsky, Greenhill, et al., 2018). If a fraction of the electorate has limited capacity to identify pieces of false information, these voters' beliefs about the candidates' competences will be manipulated so that the selection of suitable candidates for office may be severely impeded. At the same time, heterogeneity in the ability to process information can also lead to a polarization of political attitudes, which may be detrimental for society as it may increase political gridlock (Binder, 1999; Jones, 2001), decrease the provision of public goods (Alesina, Baqir, and Easterly, 1999) and aggravate the intensity of social conflict (Esteban and Ray, 1999, 2011). In order to deepen our understanding

* Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866. Financial support by the DFG, grant KR 2077/3-1, is gratefully acknowledged. Declarations of interests: none.

of how information manipulation affects our political system, this paper addresses three main questions: (i) To what extent does information manipulation by political actors impede the electorate's capability to select suitable candidates for office? (ii) How is the polarization of political attitudes in the electorate affected by false information during political campaigns? (iii) Under which conditions can policy measures that equalize the access of political candidates to the media prevent potentially detrimental effects of information manipulation?

When studying the effects of information manipulation on welfare, it is necessary to take a stance on why false stories have the power to systematically shift voting outcomes, i.e., why incorrect information is not, at least in the aggregate, entirely discounted by the electorate. In this paper, we adopt the notion that a fraction of the electorate is not fully sophisticated when processing information. In particular, we assume that these voters believe that public signals correctly represent the competence of candidates for holding office, i.e., the voters take public signals at face value. Our assumption of such naive voters captures in reduced form the common observation that individuals believe in newly arriving information even if political candidates or particular interest groups have clear incentives to manipulate them (Oliver and Wood, 2014; Allcott and Gentzkow, 2017).[1]

To explore the welfare effects of information manipulation during political elections, we set up a model that combines signal jamming (e.g., Holmström, 1999) with probabilistic voting (e.g., Enelow and Hinich, 1982; Ledyard, 1984; Lindbeck and Weibull, 1987): We consider two political candidates that compete for one term in office. The candidates are heterogeneous in their competence to be successful in office. While the exact competences of the candidates are unknown, the electorate and the candidates hold common priors about them. Voters receive public signals about the candidates' competences, which summarize their information from various sources such as television shows, radio, and social media. We assume that candidates can

1. The assumption of naive or uninformed voters is not unusual in modeling electoral competition (see, e.g., Baron, 1994; Grossman and Helpman, 1996, 2019). Moreover, this assumption is in line with stylized facts on voting behavior (see Gottfried and Shearer, 2016; Allcott and Gentzkow, 2017).

manipulate the signals by taking costly actions. For example, candidates could hire a blogger, who posts favorable stories about them on social media websites, they could discredit their opponent based on faked evidence, or hire experts to support their point of view in the media. The associated costs of spreading such false information comprise not only monetary costs but also potential future losses of reputation or even legal prosecution. While candidates' actions will distort the signals observed by the electorate, we assume, that manipulation itself is a hidden action. After having observed the public signals, voters cast their ballot. When deciding on their vote, the electorate is heterogeneous in two respects. First, voters differ in their political ideologies in the sense that every voter has a personal preference for one candidate. Second, some voters are naive as they take the observed signals at face value.

To address question (i), we compare the quality of candidate selection in equilibrium with a situation in which candidates cannot engage in manipulating voters' information. Although public signals are not manipulated in the latter situation, naive voters will nevertheless infer wrong posteriors from the available information. In particular, naive voters ignore their prior information and take signals at face value. This mistake will give an unwarranted advantage to the initially trailing candidate (i.e., the candidate with the lower prior expected competence), implying that his likelihood of winning the election is too high from a welfare perspective. Whether information manipulation further deteriorates the quality of candidate selection, depends on which candidate invests more in manipulation. We show that, if candidates receive utility from their legacy to the public, e.g., because of a higher likelihood being reelected, the disadvantaged candidate (i.e., the one with the higher prior expected competence) will invest more in information manipulation than his opponent. This result arises because it will be a particularly strong signal about his competence if he wins the election despite his handicap. The opportunity to generate such a strong competence signal creates an additional incentive for him to manipulate information. Hence, our analysis reveals a natural argument for why information manipulation can lead to a *better* quality of candidate selection.

This argument crucially relies on the assumption that candidates have homogeneous costs for spreading false information. It is, however, frequently the case that

one of the candidates has privileged access to the media because he is the incumbent and can influence part of the media, or he even owns a media station himself. We show that in situations in which one of the candidates has a sufficiently strong cost advantage in creating false information, this candidate will engage in substantially more intense manipulation than his opponent. Consequently, information manipulation can be detrimental to the quality of candidate selection. The tension between the legacy motive and the costs will also inform our discussion on policy measures to counteract the potentially adverse effect of information manipulation on welfare.

Concerning question (ii), we find that information manipulation unambiguously increases the polarization of political attitudes, which in turn harms welfare. Even without manipulation, the electorate is polarized as the beliefs of naive voters and those that correctly account for false information about candidates' competences diverge in expectation. If information manipulation becomes feasible, candidates generically choose different manipulation intensities. The additional heterogeneity leads to an aggravation of polarization in the electorate. This negative effect will be more severe the larger the difference between the candidates' manipulation intensities in equilibrium. To sum up, the welfare effect of information manipulation is positive if two requirements are fulfilled: First, candidates' costs to spread false information are similar so that the strong legacy motive of the initially more promising candidate is decisive for the candidates' manipulation incentives and, second, the quality of candidate selection is sufficiently important for welfare relative to political polarization. Otherwise, information manipulation is detrimental to welfare.

Question (iii) refers to the regulation of candidates' media access during political campaigning. Our model shows under which conditions policy measures that equalize the access of political candidates to the media are welfare enhancing in the presence of information manipulation. While in some elections politicians seem to have an almost identical media presence, in others one politician enjoys a more favorable or extensive media coverage. E.g., one of the candidates might own a media

station, can control part of the media, because he is the incumbent,[2] or has easier access to funds for his campaign. To what extent candidates are heterogeneous in their media access is often at the discretion of institutions that can restrict ownership of media companies or ensure equal funding of candidates. In our model, the media access of candidates is reflected in their costs of manipulating voters' information.

Our equilibrium results speak to the usefulness of such institutions by discussing how changes in the heterogeneity of candidates' costs of manipulating information affect welfare. A reduction in cost heterogeneity attenuates differences in manipulation intensities. Therefore, the polarization of political attitudes decreases, which enhances welfare. However, more homogeneous manipulation intensities might reduce the potential of false information to counteract the unwarranted advantage of the initially trailing candidate when voters cast their ballot. We show that, whenever costs are sufficiently homogeneous to start with, a reduction in cost heterogeneity triggers a trade-off between the quality of candidate selection and the polarization of political attitudes. The net effect of aligning candidates' media access then depends on whether welfare is mainly determined by the electorate's need for a suitable politician in office or by political polarization. However, this trade-off disappears whenever costs are sufficiently heterogeneous to start with. In other words, if one of the candidates initially has a sufficiently strong advantage in media coverage, more cost homogeneity will unambiguously increase welfare.

Related Literature

Our paper complements the literature on propaganda originated from dictators to sustain their power (see Little, 2012; Gehlbach and Simpser, 2015; Little, 2017) by studying the impact of false information on democratic elections and welfare. We adopt two critical assumptions from this literature. First, politicians can jam voters' signals in their favor by choosing unobservable actions. Second, a fraction

2. Based on field data from Hungary, Szeidl and Szucs (2017) find media capture by the government through favor exchange. The government supports specific media outlets via advertising expenditures by state-owned firms. In return, incumbent politicians receive favorable coverage by these media outlets.

of voters takes all public signals at face value. The literature on propaganda has shown that authoritarian regimes might manipulate election outcomes to secure the ongoing support of their bureaucrats (Gehlbach and Simpser, 2015) and that coordination motives within a society can induce voters to behave as if they believed the propaganda even if they do not (Little, 2017). Our model involves *two* politicians that simultaneously decide on information manipulation.[3] This setup enables us to study the effect of false information during a democratic election. In particular, we can meaningfully analyze how the manipulation of voters' information affects the quality of candidate selection and the degree of political polarization in this context.

In models of propaganda, the information distortions originate from the political actor himself. A related area of research considers how politicians choose their platforms to cater to special interest groups which then spread or influence the information in the politicians' favor (Baron, 1994; Grossman and Helpman, 1996; Bardhan and Mookherjee, 2000; Grossman and Helpman, 2001; Prat, 2002; Bardhan and Mookherjee, 2005; Medina, 2019). As in our setting these papers typically allow for a fraction of voters to be credulous in the sense that they are responsive to lobbying expenditures.[4] Politicians then have an incentive to cater to the special interest groups that raise funds for their campaigns and, thereby, distort the selection of candidates. If the fraction of credulous voters is sufficiently large, the urge to raise funds may even lead to a polarization of political platforms (Baron, 1994). We contribute to this insight by showing that false information can polarize not only political platforms but also political attitudes. Moreover, we show that the existence of special interest groups is not necessary for political polarization or the distortion of candidate selection by information manipulation.

The interplay of information manipulation, in the form of fake news, and political polarization is addressed empirically by Allcott and Gentzkow (2017) and

3. We share the assumption that two players can simultaneously distort information with Grunewald and Kräkel (2018). Their paper, however, does not study probabilistic voting and assumes that individuals correctly anticipate the amounts of false information chosen in equilibrium. A meaningful analysis of how false information affects welfare within a democratic election process is thus not possible in their framework.

4. A notable exception is Medina (2019) which assumes fully rational voters.

Lazer et al. (2018). Both papers conjecture a positive relationship between polarization and the spreading of fake news. They argue that voters with polarized beliefs are prone to believe in and disseminate positive news about their favorite political candidate while believing in negative news about political opponents. Thus, higher polarization may lead to more spreading of fake news through social media. Our model suggests that there may also be a causal relationship in the reversed direction: fake news induces political polarization. While evidence on the direction of causality is still scarce, Allcott, Braghieri, Eichmeyer, and Gentzkow (2020) show that a reduction in the usage of social media also reduces political polarization. Given our conjecture, political polarization and the spreading of fake news may reinforce each other, potentially leading to a downward spiral that can threaten social cohesion.

An important prerequisite for our analysis is that information spread through the media, correct or incorrect, affects voters' beliefs and their political attitudes. There is an ongoing discussion to what extent media affect preferences. Gerber, Karlan, and Bergan (2009) and Gentzkow, Shapiro, and Sinkinson (2011) show that biased reports in newspapers do not seem to influence political opinions. In contrast, those studies that deal with biased reports that are circulated via modern mass media have documented an impact on voter preferences in US presidential and gubernatorial elections (DellaVigna and Kaplan, 2007; Gerber, Gimper, Green, and Shaw, 2011), the Weimar Republic (Adena, Enikolopov, Petrova, Santarosa, and Zhuravskaya, 2015), and parliamentary as well as presidential elections in the Ukraine (Peisakhin and Rozenas, 2018). These findings, in turn, have immediate implications for the policy tools that we analyze. In particular, we show that the detrimental effects of information manipulation on welfare may be attenuated if no candidate enjoys favorable access to the media. Combining this result with the existing insights suggests that homogeneous access to media is particularly important for mass media like television, radio, and social media rather than for print media.

Most closely related to our analysis is the paper by Grossman and Helpman (2019) on the effect of information manipulation in electoral competitions. They also make use of a probabilistic voting model but build upon the premise that information manipulation by politicians is costless. In this setup, they focus on the

polarization of party platforms and show that information manipulation can lead to more polarization therein. We complement their results by analyzing candidate selection and the polarization of political attitudes in the electorate. While their model predicts that less information manipulation reduces polarization on the party level, we show that less information manipulation may induce a trade-off between worse candidate selection and reduced polarization of political attitudes. Given this trade-off, we can discuss to what extent policy measures that equalize candidates' access to the media can help to counteract the adverse effects of information manipulation.

1.2 Model

In this section, we propose a model to analyze how information manipulation in electoral competitions affects voters' behavior and welfare. For this purpose, we study an electoral setting with a continuum of heterogeneous voters and two candidates, who can manipulate voters' beliefs to boost their chances of winning the election.

The Candidates

There are two candidates, A and B, who have unknown competences for being successful in office.[5] We denote candidate j's ($j = A, B$) competence by q_j. While the exact competence of each candidate is unknown to candidates and voters ex ante, their distributions $q_j \sim N\left(\bar{q}_{j0}, \sigma_0^2\right)$ with $\bar{q}_{j0} \geq 0$ for $j \in \{A, B\}$ are common knowledge.[6] For example, we can imagine that there is common uncertainty about the future and that neither the heterogeneous candidates nor the electorate know which candidate can better cope with the challenges of the upcoming political term.[7] We assume that candidate B is ex ante associated with a higher expected competence, i.e., $\bar{q}_{B0} > \bar{q}_{A0}$.

5. For a discussion on the importance of competence in politics and the selection of appropriate politicians for office see Besley (2005).

6. Symmetric uncertainty is a common assumption in the signal-jamming literature; see, e.g., Stein (1989), Meyer and Vickers (1997), Holmström (1999), Höffler and Sliwka (2003), Bar-Isaac and Deb (2014), and Miklós-Thal and Ullrich (2014).

7. Our model is also analytically equivalent to a model in which candidates' competences are heterogeneous and commonly known, but there exists symmetric uncertainty about an underlying state of the world that determines whose candidate's competences are more valuable for the electorate during the upcoming political term.

Before the election, voters receive a public and informative signal for each candidate (e.g., from newspapers, television shows, and online forums). While we assume these two signals to reveal some information on the true underlying competences, candidates can also manipulate them. Candidate j could, for example, spread deliberate misinformation via social media or traditional print. Let $m_j \in [0, \bar{m}]$, with $\bar{m} > 0$ being finite, denote the manipulation intensity chosen by candidate j. Crucially, these intensities are hidden actions, and the manipulation of signals can neither be directly recognized nor eliminated by a candidate's opponent or the voters.[8]

Manipulating information comes at some costs to political candidates such as monetary costs of hiring bloggers or agencies, a potentially increased likelihood of legal prosecution, and potential compensation payments after the election. We represent the respective costs by the functions $c_j\left(m_j\right) = \kappa_j \cdot c\left(m_j\right)$ with $\kappa_j > 0$ $(j = A, B)$, $c\left(0\right) = c'\left(0\right) = 0$, $c'\left(m_j\right), c''\left(m_j\right) > 0$ for $m_j > 0$, and $\lim_{m\to\infty} c'\left(m\right) = \infty$ $(j = A, B)$. Moreover, we assume that $c''(m)$ is bounded from below with $c''(m) \geq \underline{c}$. Candidates' cost parameters κ_j capture potential differences in their ability of manipulating beliefs. Such heterogeneity might stem from differences in available campaigning budgets as well as differences in candidates' ability to manipulate media content if, e.g., one of the candidates owns a media company and can implement media censorship.

Given manipulation intensities and the candidates' competences, the public signals that are observed by the electorate take the following form:[9]

$$s_j \;=\; q_j + m_j + \epsilon_j \qquad\qquad (j \;=\; A, B), \tag{1.1}$$

where the noise terms ϵ_j follow a standard normal distribution also comprise the influence of exogenous misinformation, which is not deliberately spread by the two

8. We abstract from endogenous platform choices by the politicians. However, our results carry over to a model in which candidates have fixed platforms, e.g., due to their party membership or political ideology, and information is manipulated to influence signals about an underlying state of the world, which determines the welfare consequences of the platforms in the upcoming term.

9. In line with the literature on signal jamming, we assume the resulting public signals to have an additive structure; see for example Holmström (1999) and Little (2017). To ensure readability, we focus on positive information manipulation about a candidate's own competence. However, as the electorate's relative comparison between the two candidates is decisive for the election outcome, our model also captures the case of manipulations that aim at discrediting an opponent's competence.

candidates (for examples see Allcott and Gentzkow, 2017). We assume that all random variables are stochastically independent.

Candidates are motivated to enter office for two reasons: First, as is commonly assumed in the literature, they receive personal benefits, which we denote by b. Second, we follow Maskin and Tirole (2004) and assume that politicians wish to leave a legacy, i.e., they want to be remembered for having done great things while in office. Since the legacy motive is crucial for the results to come, it deserves some more discussion.[10] A legacy motive cannot only arise if politicians are interested in the quality of the policies that they implement but also if they care about how the public perceives their competence. Suppose, for example, each politician cares about being reelected, and the electorate will receive an informative signal about her competence during her term. In this case, politicians with a higher competence expect the signal to be more favorable than politicians with a lower competence. If voters base their future voting decisions on this signal, politicians' utilities of holding office increase with their actual competence – they have a legacy motive. We conceptualize these motives by assuming that a candidates' utilities is given by:[11]

$$v_j\left(m_j\right) = \begin{cases} b + q_j - c_j(m_j) & \text{if } j \text{ wins the election} \\ -c_j(m_j) & \text{otherwise.} \end{cases}$$

We assume candidates to be sophisticated in the sense that they form rational beliefs about their opponent's behavior and the behavior of all voters.

The Electorate

There is a unit mass of risk-neutral heterogeneous voters who determine the winner of the election by simple majority rule. Before voters cast their ballot, they make inferences from the received public signals on the candidates' competences. To model

10. In a model in which candidates are not motivated by their legacy, it turns out that the equilibrium is symmetric in the sense that both candidates choose the same manipulation intensity. As voters care only about the relative competence of candidates, in a symmetric equilibrium the outcome of the election is not affected by information manipulation.

11. To keep the model simple, we assume legacy and office motive to be additively separable. All results also transfer to a setting in which utility in case of wining is given by $bq_j - c_j(m_j)$, or $b(q_j - c_j(m_j))$.

the impact of information manipulation in the political process, we embrace the notion that manipulation is effective because a fraction α of voters is *naive*. These voters hold the belief that public signals correctly represent the quality of candidates, i.e., they take the public signals at face value.[12] Their posterior expectation about the quality of candidate j is thus given by s_j for $j \in \{A, B\}$. Importantly, such voters are not able to account for potential distortions in the signals due to information manipulation by the candidates. This assumption captures in reduced form the common observation that individuals believe in newly arriving information even if this information seems obviously false or is generated by groups with particular interests (Oliver and Wood, 2014; Allcott and Gentzkow, 2017).[13] The remaining fraction $1-\alpha$ of the electorate is *sophisticated*. These voters anticipate that politicians can manipulate signals, and update their priors correctly.

Every voter cares about the politician in office for two reasons. First, a more competent politician will increase the overall surplus for the electorate. Second, following the literature on probabilistic voting,[14] voters have idiosyncratic preferences over candidates. We will refer to these preferences as the voters' *ideology*. Ideologies are driven by individual characteristics of the candidates that are not connected to their competence, like their party membership, their public behavior, and their outward appearance. The idiosyncratic parameter θ measures by how much a voter prefers candidate A over candidate B because of his ideology. It is commonly known that ideologies in the electorate are distributed according to G_θ, where the corresponding density g_θ is symmetric with a unique mode at zero and has full support over $\mathbb{R}$. The variance of θ is denoted by σ_θ^2. Overall, the utility of a voter with ideology θ from the outcome of the election is given by

$$u_\theta = \begin{cases} \theta + q_A & \text{if } A \text{ wins the election} \\ q_B & \text{if } B \text{ wins the election.} \end{cases}$$

12. We share this assumption with Grossman and Helpman (2019).

13. While this assumption simplifies the analysis, our results do not hinge on the exact specification of voter naivety. In the conclusions, we discuss alternative forms of voter naivety under which our qualitative results also hold.

14. See, e.g., Enelow and Hinich (1982), Ledyard (1984) and Lindbeck and Weibull (1987).

As voters are heterogeneous in their sophistication and their ideologies, they will hold different views on the candidates' qualifications to hold office. We define the *political attitude* of a voter with ideology θ and sophistication type $i \in \{S,N\}$ by

$$\Delta u^i_\theta \;:=\; \theta + \bar{q}^i_{A1} - \bar{q}^i_{B1}, \tag{1.2}$$

where $\bar{q}^i_{j1}$ with $i \in \{S,N\}$ denotes a voter's posterior expectation about candidate j's competence, and the superscripts "S" and "N" indicate a sophisticated and a naive voter, respectively. His political attitude is hence described by how much he favors candidate A over candidate B after the realization of signals.

In sum, the game evolves according to the following three steps. First, candidate $j \in \{A,B\}$ chooses manipulation intensity m_j and incurs costs $c_j\left(m_j\right)$. Second, public signals realize, and voters update their prior beliefs. Third, voters cast their ballot, and the winner of the election is chosen via simple majority rule.

Solution Concept

We apply a slightly modified version of pure-strategy Bayesian Nash Equilibrium to account for the fact that some voters are naive. Moreover, we assume that voters vote sincerely, i.e., they decide based on their perceived posterior expectation about the candidates' competences and their ideology.[15] An equilibrium of the game then consists of a pure-strategy profile incorporating pure strategies of both candidates and all voters, as well as a belief system such that the following two statements hold. First, both candidates play mutually best responses conditional on voters voting sincerely. Second, a fraction α of voters is naive and believes that signals correspond to true underlying qualities, whereas sophisticated voters correctly anticipate manipulation intensities along the equilibrium path.[16]

15. In our model, being the pivotal voter does not contain any information for sophisticated voters. Hence, assuming sincere voting is without loss of generality for this group of voters. For naive voters being pivotal might, however, contain information about their own misperception. Their strategic voting behavior is equivalent to sincere voting if they are sure that nobody else is more capable in processing information than they are.

16. This equilibrium concept is a Berk-Nash equilibrium (see Esponda and Pouzo, 2016) as it allows some players to believe in a misspecified model of the world on the equilibrium path. In our setting, naive voters believe in the particularly restrictive model of the world that information manip-

1.3 Equilibrium Analysis

We solve for the equilibrium of the game by first analyzing the voters' beliefs and behavior. Afterward, we derive the candidates' optimal manipulation intensities. Making use of the insights on updating of normally distributed beliefs by DeGroot (2005) immediately leads to the following results: If sophisticated voters hold a degenerate belief, $\hat{m}_j$, about candidate j's hidden manipulation intensity, their posterior mean of candidate j's competence ($j = A, B$) will be given by

$$\bar{q}_{j1}^S \;=\; \bar{q}_{j0} + \Sigma \cdot \left(s_j - \bar{q}_{j0} - \hat{m}_j\right), \text{ with } \Sigma \;:=\; \frac{\sigma_0^2}{\sigma_0^2 + 1}. \tag{1.3}$$

According to equation (1.3), a sophisticated voter will improve his opinion about candidate j only if the acquired information on that candidate is favorable in light of the anticipated amount of manipulation ($s_j > \bar{q}_{j0} + \hat{m}_j$). The term Σ denotes the weight that voters place on newly arriving information compared to their prior expectations.[17] The higher the initial uncertainty about the candidate's competence (i.e., σ_0^2 is large), the more the sophisticated voters will rely on new information. In contrast, naive voters' posterior mean of candidate j's competence will be given by

$$\bar{q}_{j1}^N \;=\; s_j. \tag{1.4}$$

Naive voters will, therefore, be systematically fooled about the actual competence of candidates in equilibrium because they do not account for the amount of manipulation chosen by politicians.[18]

Given the voters' updating behavior, the political attitude of a sophisticated voter with ideology θ is

$$\Delta u_\theta^S \;=\; \theta + (1 - \Sigma)\left(\bar{q}_{A0} - \bar{q}_{B0}\right) - \Sigma\left(m_B - \hat{m}_B - m_A + \hat{m}_A + \gamma\right), \tag{1.5}$$

---

ulation does not exist. Note, however, that candidates in our setting have a continuous action space, whereas all players in Esponda and Pouzo (2016) have finite action spaces such that we cannot immediately apply their existence result.

17. Our statement only specifies the updating process for degenerate beliefs about the candidates' hidden manipulation intensities. As we consider equilibria in pure strategies and sophisticates form correct beliefs in equilibrium, this characterization will suffice for our later analysis.

18. A qualitatively equivalent approach would be to follow Little (2019), and to assume that naive voters choose their beliefs to balance the tension between plausible conclusions from the signals and those beliefs that they wish to hold.

and the political attitude of a naive voter with ideology θ is

$$\Delta u_\theta^N = \theta - \big(m_B - m_A\big) - \gamma \tag{1.6}$$

with $\gamma := q_B - q_A + \epsilon_B - \epsilon_A$. The random variable γ summarizes all stochastic elements that are relevant for the behavior of voters. The composed random variable is larger the more favorable the acquired information about candidate B is relative to that about candidate A. We denote the cdf of the random variable γ by G_γ and the corresponding density by g_γ. As can be seen from its definition, γ is normally distributed with mean $\mu_\gamma = \bar{q}_{B0} - \bar{q}_{A0} > 0$. Recall that a voter's political attitude measures how much he prefers candidate A over candidate B. As γ will increase if signals contain more favorable information about candidate B compared to candidate A, a voter's political attitude is a decreasing function of the realization of γ.

Political attitudes determine whom the voters cast their ballot for. Since we assumed sincere voting, a voter with ideology θ and sophistication type $i \in \{S,N\}$ will cast his ballot for candidate A if and only if $\Delta u_\theta^i > 0$. As the political attitude of any voter decreases in the realization of γ, the fraction of voters that vote for candidate A increases if the realization of γ decreases. Consequently, there exists a cutoff $\tilde{\gamma}$ such that each candidate receives exactly 50 percent of the votes if and only if $\gamma = \tilde{\gamma}$, being implicitly defined by

$$\begin{aligned}\frac{1}{2} =& \alpha \cdot G_\theta(m_B - m_A + \tilde{\gamma}) \\ &+ (1-\alpha) \cdot G_\theta((1-\Sigma)\mu_\gamma + \Sigma\big(m_B - \hat{m}_B - m_A + \hat{m}_A\big) + \Sigma\tilde{\gamma}).\end{aligned} \tag{1.7}$$

As candidate A is sophisticated and will win the election if and only if $\gamma \leq \tilde{\gamma}$, his objective function is

$$G_\gamma(\tilde{\gamma}) \cdot E\big[b + \bar{q}_{A1}^S \big| \gamma \leq \tilde{\gamma}\big] - c_A(m_A),$$

and candidate B's objective function is

$$\big(1 - G_\gamma(\tilde{\gamma})\big) \cdot E\big[b + \bar{q}_{B1}^S \big| \gamma > \tilde{\gamma}\big] - c_B(m_B).$$

Conditional on being elected, each candidate benefits from a high utility of holding office, b, and his legacy. As candidates update correctly, their expectations about

their competence will equal those of sophisticated voters. The first line of equation (1.7) shows that, ex ante, the presence of naive voters reduces the winning chances of the initially leading candidate B – the naifs ignore his prior lead, μ_γ, and, therefore, tend to vote for the trailing candidate too often (compare the arguments of $G_\theta(\cdot)$ in the first and the second line of equation (1.7)). Consequently candidate B has to pass a higher cutoff $\tilde{\gamma}$ for winning the election compared to a situation without naive voters. Concerning the candidates' objective functions, this handicap boosts candidate B's manipulation incentives: Because of his handicap, candidate B's posterior expected legacy conditional on winning the election will be particularly large. We obtain the following equilibrium result:[19]

Proposition 1.1. *If $\underline{c}$ and $\bar{m}$ are sufficiently large, there will exist an equilibrium in pure strategies. The equilibrium is interior and candidates' manipulation intensities are described by*

$$g_\gamma(\tilde{\gamma}^*)b + g_\gamma(\tilde{\gamma}^*)\left(\bar{q}_{A0} - \frac{\Sigma}{2}(\tilde{\gamma}^* - \mu_\gamma)\right) = c_A'(m_A^*) \tag{1.8}$$

and

$$g_\gamma(\tilde{\gamma}^*)b + g_\gamma(\tilde{\gamma}^*)\left(\bar{q}_{B0} + \frac{\Sigma}{2}(\tilde{\gamma}^* - \mu_\gamma)\right) = c_B'(m_B^*) \tag{1.9}$$

with $\tilde{\gamma}^* = \tilde{\gamma}(m_A^*, m_B^*)$ satisfying

$$\frac{1}{2} = \alpha \cdot G_\theta(m_B^* - m_A^* + \tilde{\gamma}^*) + (1-\alpha) \cdot G_\theta((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^*). \tag{1.10}$$

The proposition characterizes the candidates' equilibrium behavior. Equations (1.8) – (1.10) show how the cutoff $\tilde{\gamma}^*$ and the candidates' manipulation intensities interact in equilibrium. Equation (1.10) reveals that equilibrium cutoff $\tilde{\gamma}^*$ crucially depends on the difference in manipulation intensities, $m_B^* - m_A^*$. Compared to a situation in which both candidates choose the same amount of manipulation, the equilibrium cutoff will be higher if $m_A^* > m_B^*$ and lower if $m_A^* < m_B^*$.

The two equations (1.8) and (1.9) describe the incentives to manipulate information in equilibrium. The first term on the left-hand sides of (1.8) and (1.9)

19. All proofs are relegated to the appendix.

depicts a candidate's incentives in anticipation of the potential office benefit b. It is identical for both candidates.

The second term describes the incentives of a candidate to manipulate information in anticipation of his legacy to the public. It is determined by the product of the marginal winning probability and the expected legacy of the candidate, when the random variable γ coincides with the equilibrium cutoff $\tilde{\gamma}^*$. To build up an intuition for this effect, consider a marginal increase in manipulation as a unilateral deviation by one candidate. This increase raises the candidate's probability of winning the election, as he will convince some voters to cast their ballot in his favor. As the candidate now wins an election that would have been a tie without this deviation, his additional utility is given by the marginal gain in his winning probability, $g_\gamma(\tilde{\gamma}^*)$, times his legacy in the event of a tie, i.e. when $\gamma = \tilde{\gamma}^*$. In equilibrium, these considerations generically lead to heterogeneous incentives for the candidates, because they might expect different legacies conditional on a tied election.

1.4 Welfare

In our analysis, we consider ex-ante expected welfare, i.e., welfare before signals and candidates' competences are realized. We build on the premise that candidate selection and the polarization of political attitudes in the electorate are integral parts of welfare. Therefore, our welfare measure consists of two components. First, every voter receives utility from the elected candidate's administration, which depends on the politician's competence and the voter's ideology. Second, welfare decreases with the polarization of political attitudes within the electorate. Such polarization can be detrimental for society as it may increase political gridlock (Binder, 1999; Jones, 2001), aggravate the intensity of social conflict (Esteban and Ray, 1999, 2011), and lead to less public good provision (Alesina, Baqir, and Easterly, 1999). Following DiMaggio, Evans, and Bryson (1996), we consider the ex-ante expected variance of political attitudes as our measure of polarization. In Appendix 1.B, we derive a micro foundation for why polarization can be detrimental to welfare. For this purpose, we build upon the idea that differences in political attitudes may aggravate social

conflict. In line with that idea our micro foundation is based on a variation of the models established by Esteban and Ray (1999) and Esteban and Ray (2011).

Let $h^i(\Delta u^i_\theta|\gamma)$ denote the density of political attitudes for voters of type $i \in \{N,S\}$ given the realization of γ. Then we can define the density of political attitude Δu in the electorate as $h(\Delta u|\gamma)$ by

$$h\left(\Delta u|\gamma\right) = \alpha h^N\left(\Delta u|\gamma\right) + (1-\alpha)\, h^S\left(\Delta u|\gamma\right).$$

As we take an ex ante perspective, Δu is a random variable whose realizations depend on q_j and ϵ_j with $j \in \{A,B\}$. Let $\sigma^2_{\Delta u}$ denote the variance of Δu. Then, we can define our welfare measure as follows:

Definition 1.1. *Welfare $\mathscr{W}$ is defined as*

$$\mathscr{W} := E[u_\theta] - \beta \cdot E\left[\sigma^2_{\Delta u}\right], \tag{1.11}$$

where $\beta \geq 0$ determines the importance of political polarization relative to candidate selection. The expectation operator $E[\cdot]$ is taken with respect to q_A, q_B, and the random variables ϵ_A, ϵ_B.

In our welfare measure, $E[u_\theta]$ and $\beta E\left[\sigma^2_{\Delta u}\right]$ reflect how welfare is affected by candidate selection and political polarization, respectively. The parameter β allows for an arbitrary weighting of the two components. It measures how strong societal well-being is impaired by political gridlock, social conflict and missing cooperation in providing public goods. Given our insights from the previous section, we can derive the following characterization of welfare:

Proposition 1.2. *In equilibrium, welfare $\mathscr{W}$ amounts to*

$$\mathscr{W}^* = \mathscr{E}(\tilde{\gamma}^*) - \beta \cdot \mathscr{P}\left(m^*_A, m^*_B\right) \tag{1.12}$$

with

$$\mathscr{E}(\tilde{\gamma}^*) := E[q_A|\gamma < \tilde{\gamma}^*] \cdot G_\gamma\left(\tilde{\gamma}^*\right) + E[q_B|\gamma \geq \tilde{\gamma}^*] \cdot \left(1 - G_\gamma\left(\tilde{\gamma}^*\right)\right) \tag{1.13}$$

and

$$\mathscr{P}\left(m^*_A, m^*_B\right) := \sigma^2_\theta + \alpha(1-\alpha)\left[(m^*_A - m^*_B)^2 + (1-\Sigma)^2\sigma^2_\gamma\right]. \tag{1.14}$$

The first part of welfare, $\mathscr{E}(\tilde{\gamma}^*)$, reflects the expected utility that a random voter receives from the elected candidate's administration in equilibrium. Recall that a random voter in our setting is ideologically unbiased. Hence, his expected utility from the administration is identical to the candidate's expected competence conditional on winning the election, which yields expression (1.13). This component of welfare is pinned down by the location of the equilibrium cutoff $\tilde{\gamma}^*$.

The second part of welfare, $\mathscr{P}\left(m_A^*, m_B^*\right)$, is determined by the variance in political attitudes across the population. Whether or not political attitudes will be similar in expectations does not only depend on the distribution of ideologies but also on the distribution of voters' beliefs about the competences of political candidates. A fundamental force that is driving polarization in our model is thus the divergence of political attitudes of naive and sophisticated voters due to the manipulation of information by the candidates during their campaigns. Finally, polarization is also influenced by the group sizes of the two types of voters.

Equation (1.14) implies immediately that the manipulation of information – e.g., via fake news – (weakly) increases political polarization. This result is an important complement to earlier arguments. Allcott and Gentzkow (2017) and Lazer et al. (2018) argue that polarization can be the perfect breeding ground for the spreading of fake news because polarized voters are keen on hearing positive news about their favorite candidate and negative news about their favorite's opponent. Therefore, higher polarization leads to more fake news. Our findings support the idea that the direction of causality may also be reversed, i.e., fake news exacerbates political polarization. Jointly, both views imply that fake news and political polarization might aggravate each other leading to a downward spiral with detrimental consequences for social cohesion.

Benchmark

As a benchmark case for our welfare analysis, we consider a situation in which there is *no manipulation* of voters' information (*NM*). In this situation, the cutoff $\tilde{\gamma}^{NM}$ can be derived from (1.7) by acknowledging that $m_j = \hat{m}_j = 0$ $(j = A, B)$. Therefore, it is given by

$$\frac{1}{2} = \alpha \cdot G_\theta(\tilde{\gamma}^{NM}) + (1-\alpha) \cdot G_\theta((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^{NM}). \tag{1.15}$$

If politicians cannot engage in information manipulation, the amount of political polarization can be derived from (1.14) and is given by

$$\mathscr{P}^{NM} := \sigma_\theta^2 + \alpha(1-\alpha)(1-\Sigma)^2\sigma_\gamma^2. \tag{1.16}$$

As illustrated by Equation (1.16), political polarization is present even in this benchmark case and emerges from two sources. First, voters differ in their ideology θ. Second, voters also differ in how they form their posterior expectations about candidates' competences. As naive voters take public signals at face value, they will disagree with sophisticates about the quality of the political candidates. Overall, welfare for the case without information manipulation amounts to

$$\mathscr{W}^{NM} := \mathscr{E}\left(\tilde{\gamma}^{NM}\right) - \beta \cdot \left(\sigma_\theta^2 + \alpha(1-\alpha)(1-\Sigma)^2\sigma_\gamma^2\right). \tag{1.17}$$

In the following, we will compare the welfare that arises in equilibrium, $\mathscr{W}^*$, with our benchmark welfare, $\mathscr{W}^{NM}$.

1.5 Welfare Effects of Information Manipulation

As naive voters take signals at face value, even in the absence of information manipulation they are excessively responsive to newly arriving information compared to sophisticated voters. Such excessive responsiveness favors the candidate that is initially trailing because differences in prior competence beliefs are ignored by the naifs and, hence, become less predictive for the election outcome. If information manipulation is not feasible, the electoral outcome will, therefore, be distorted in favor of candidate A, whose prior expected competence is lower ($\bar{q}_{B0} > \bar{q}_{A0}$). More formally, the no-manipulation cutoff, $\tilde{\gamma}^{NM}$, will be higher than the cutoff that would arise if all voters were sophisticated. As the average sophisticate is ideologically unbiased, an electorate that only consists of sophisticated voters elects the welfare maximizing candidate. The corresponding *sophisticated-voting cutoff*, $\tilde{\gamma}^{SV}$, can be derived by setting $\alpha = 0$ and $m_j = \hat{m}_j$ ($j = A, B$) in (1.7):

$$\frac{1}{2} = G_\theta\left((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^{SV}\right) \Leftrightarrow \tilde{\gamma}^{SV} = -\frac{1-\Sigma}{\Sigma}\mu_\gamma. \tag{1.18}$$

Comparing the no-manipulation cutoff with the sophisticated-voting cutoff leads to the following result:

Lemma 1.1. *$\tilde{\gamma}^{NM} > \tilde{\gamma}^{SV}$, implying $\mathscr{E}(\tilde{\gamma}^{NM}) < \mathscr{E}(\tilde{\gamma}^{SV})$ with $\partial \mathscr{E}(\tilde{\gamma}^{NM})/\partial \alpha < 0$.*

Thus, compared to $\tilde{\gamma}^{SV}$, the composed random variable γ has to pass a higher cutoff $\tilde{\gamma}^{NM}$ for candidate B to win the election. In other words, according to Lemma 1.1, the election outcome will be distorted in favor of candidate A if a subgroup of voters suffers from naivety and information manipulation is not feasible. This deterioration of candidate selection will be reinforced by a larger fraction of naive voters in the electorate. As naivety also drives apart the political attitudes of naive and sophisticated voters (see (1.16)), polarization increases by the amount $\alpha(1-\alpha)(1-\Sigma)^2\sigma_\gamma^2$ compared to a situation without naivety (i.e., $\alpha = 0$). This increase in polarization will be larger the more equal the shares of naive and sophisticated voters. Overall, naivety will thus unambiguously reduce welfare in the absence of information manipulation. In the following, we analyze whether the welfare decreasing effect of naivety will be aggravated if the candidates can additionally manipulate the voters' information. We first consider the case of homogeneous costs.

Proposition 1.3. *If $\kappa_A = \kappa_B$, then $m_B^* > m_A^*$, and $\exists\ \beta^* > 0$ such that $\mathscr{W}^* \geq \mathscr{W}^{NM}$ if $\beta \leq \beta^*$ and $\mathscr{W}^* < \mathscr{W}^{NM}$ if $\beta > \beta^*$.*

If costs are homogeneous across candidates, candidate B will more intensely manipulate information than candidate A. This result originates from the diverging strengths of both candidates' legacy motives. Recall that naivety favors the initially trailing candidate A. If in this situation candidate B wins the election, it is a particularly strong signal about B's competence. Consequently candidate B's legacy motive will be stronger than A's, which induces candidate B to choose a higher manipulation intensity. Therefore, a part of A's unwarranted advantage due to naivety is eliminated: Information manipulation serves as a debiasing device, and the election outcome improves relative to the situation with no manipulation ($\mathscr{E}(\tilde{\gamma}^*) > \mathscr{E}(\tilde{\gamma}^{NM})$).

However, this debiasing effect of information manipulation comes at a price. As the candidates choose different manipulation intensities in equilibrium, the degree of polarization increases by $\alpha(1-\alpha)(m_A^* - m_B^*)^2$ compared to the situation

without manipulation (cp. equation (1.14)). In particular, the asymmetric choice of manipulation intensities additionally disperses political attitudes and, hence, aggravates polarization, which reduces welfare. All in all, information manipulation triggers a trade-off between aggravating polarization and improving election outcomes. Whether the net welfare effect of information manipulation is positive or not is thus determined by the magnitude of β in (1.12).

At first glance, it is surprising that the welfare effect of information manipulation depends on the value of β. In particular, the debiasing effect of information manipulation seemingly aligns the beliefs of naive and sophisticated voters. Consequently, one would expect a decrease in polarization whenever candidate selection improves. Crucially, however, information manipulation is only pivotal for the electoral outcome if the realization of γ implies that candidate A would win the election by a small margin if there were no manipulation of information. In these situations, information manipulation counteracts the mistakes of naive voters. In contrast, for other realizations of γ information manipulation is not pivotal for the electoral outcome but drives apart political attitudes of sophisticated and naive voters. In expectation, it is, therefore, detrimental to welfare via increasing polarization in the electorate even if it improves candidate selection.

While the trade-off between an increase in polarization and a better candidate selection is particularly apparent in the case of homogeneous costs, it is not specific to this case. The manipulation of information may improve welfare even under heterogeneous costs. Let $\Delta\kappa := (|\kappa_A - \kappa_B|)/(\kappa_A + \kappa_B)$. Then the following proposition shows a sufficient condition for the trade-off to arise if costs are heterogeneous.

Proposition 1.4. *Suppose that* $\kappa_B < \kappa_A$. *If* $\alpha \geq \frac{1}{2}$ *and*

$$\bar{q}_{B0}\left(\frac{1-\Sigma}{1+\Sigma} - \Delta\kappa\right) - \bar{q}_{A0}\left(\frac{1-\Sigma}{1+\Sigma} + \Delta\kappa\right) \geq 2b \cdot \Delta\kappa, \tag{1.19}$$

then $\exists\ \beta^{**} > 0$ *such that* $\mathscr{W}^* \geq \mathscr{W}^{NM}$ *if* $\beta \leq \beta^{**}$ *and* $\mathscr{W}^* < \mathscr{W}^{NM}$ *if* $\beta > \beta^{**}$.

Proposition 1.4 shows that, in general, two conditions have to be met for information manipulation to improve candidate selection. First, it is crucial that candidate B has a cost advantage so that he manipulates more than A. This condition

ensures that information manipulation counteracts the advantage that candidate A enjoys due to naivety. Second, the distortions induced by the naive voters' neglect of their prior information about the candidates' competences need to be large relative to those that are induced by information manipulation. Otherwise, information manipulation may eliminate the naifs' initial distortions but at the same time creates distortions that outweigh the eliminated ones. This presumption is jointly ensured by inequality (1.19) and $\alpha \geq \frac{1}{2}$: The distortions stemming from naivety will be large if the share of naive voters, α, is large, if the initial lead of candidate B, $\bar{q}_{B0} - \bar{q}_{A0}$, which is ignored by the naifs, is large, and if prior information is sufficiently important relative to new information (i.e., Σ is small). At the same time, the distortions generated by a higher manipulation intensity of candidate B are moderate if office motivation b and B's cost advantage, $\kappa_A - \kappa_B$, are small. If the two presumptions are met, the existence of information manipulation induces the same trade-off between an aggravation of polarization and an improvement of candidate selection as in the case of homogeneous costs.

While the manipulation of information may function as a debiasing device under the parameter constellations discussed above, it can clearly also be the case that such manipulation is detrimental to the quality of candidate selection. The following Proposition derives parameter constellations under which this holds.

Proposition 1.5. *Suppose that $\kappa_B > \kappa_A$. If*

$$\bar{q}_{B0}\,(1 - \Sigma - \Delta\kappa) - \bar{q}_{A0}\,(1 - \Sigma + \Delta\kappa) \;<\; 2b \cdot \Delta\kappa, \tag{1.20}$$

then $\mathscr{W}^* < \mathscr{W}^{NM}$ for all β.

Proposition 1.5 shows that there exist constellations in which the manipulation of information is unambiguously detrimental for welfare. Now, candidate A has a cost advantage over candidate B so that distortions caused by information manipulation and naivety may aggravate each other. This will be the case if candidate A invests more in information manipulation than candidate B, which holds true if the cost advantage of A is sufficiently strong so that candidate B's high incentive to manipulate information due to his higher legacy motive is overruled. Condition (1.20)

shows, which parameter constellations satisfy this requirement: The condition will be satisfied if $\kappa_B - \kappa_A$ as well as office motivation b are sufficiently large, candidate B's initial lead is small ($\bar{q}_{A0} \approx \bar{q}_{B0}$), and prior information is rather unimportant relative to new information (i.e., Σ is large). In particular, condition (1.20) holds whenever $\Delta\kappa > 1 - \Sigma$. As information manipulation always increases polarization, it then has a detrimental impact on welfare for all β.

Our results suggest that information manipulation during political campaigns is particularly detrimental for voters' welfare if the candidates' costs to spread misinformation differ substantially. Then, information manipulation not only increases political polarization but also impedes candidate selection. If costs are similar across candidates, however, information manipulation improves candidate selection while aggravating polarization of political attitudes. In the next section, we build upon this idea and analyze to what extent institutions that equalize candidates' media access can help to overcome adverse effects of information manipulation in elections.

1.6 Equalizing Candidates' Media Access

In our setup, the marginal costs of candidates reflect their capacity to manipulate information and to avoid future prosecution. In particular, candidates that can directly influence media coverage because they own media stations might have a high capacity to manipulate voters' information. To address how policies that equalize candidates' media access affect welfare, we will next provide comparative statics results on candidates' marginal costs. The implicit function theorem ensures that the manipulation intensities of both candidates are locally continuously differentiable in every equilibrium. Therefore, the comparative statics are well defined for all equilibria. Moreover, our results show that even if there exist multiple equilibria, the comparative statics have the same sign for every equilibrium.[20] The following

20. Proposition 1.7 in the appendix additionally provides sufficient conditions on the candidates' cost function under which the equilibrium is unique.

proposition shows under which conditions it is welfare improving to equalize candidates' access to the media.[21]

Proposition 1.6. *For every κ_i with $i \in \{A,B\}$, there exist unique $\underline{\kappa}_j(\kappa_i)$ and $\bar{\kappa}_j(\kappa_i)$ with $j \in \{A,B\}$, $j \neq i$ and $\underline{\kappa}_j(\kappa_i) < \kappa_i < \bar{\kappa}_j(\kappa_i)$, such that the following statements hold:*

(i) If $\kappa_j \notin \left[\underline{\kappa}_j(\kappa_i), \bar{\kappa}_j(\kappa_i)\right]$, then $\mathscr{W}^*$ will increase for all β if κ_j moves towards κ_i.

(ii) If $\kappa_j \in \left(\underline{\kappa}_j(\kappa_i), \bar{\kappa}_j(\kappa_i)\right)$, then $\partial \mathscr{W}^* / \partial \kappa_j > 0$ if $j = A$ and β is sufficiently small, or if $j = B$ and β is sufficiently large.

Case (i) of Proposition 1.6 deals with parameter constellations where cost heterogeneity is large. In such situations, a decrease in heterogeneity will reduce polarization and increase the quality of candidate selection so that welfare clearly increases. Intuitively, a large cost heterogeneity implies that the candidate with lower costs will invest substantially more in information manipulation than his counterpart with higher costs. Consequently, he will increase his chances to win the election compared to a situation in which all voters are sophisticated. A reduction in cost heterogeneity will diminish the difference in manipulation intensities and, thereby, the unwarranted advantage that the low-cost candidate receives. Therefore, candidate selection improves. At the same time, the reduced difference in manipulation intensities also decreases polarization. Overall, political institutions should, therefore, ensure that candidates access to the media and, thus, their capacity to directly influence media coverage is either generally limited [22] or not too unequal.

Case (ii) of Proposition 1.6 shows that both welfare components – $\mathscr{E}(\tilde{\gamma}^*)$ as well as $\mathscr{P}\left(m_A^*, m_B^*\right)$ – will be differently affected if cost heterogeneity is small. If costs are similar, candidate B will manipulate more than candidate A due to his

21. While we analyze the impact of policies that reduce the heterogeneity in candidates' costs of manipulating information, it might sometimes be infeasible to impede the media access of one particular candidate only. For example, eliminating fake news on social media will affect both candidates similarly. In our setup, we can analyze such policies by assuming that there are common components to candidates' costs of manipulating information, i.e, $\kappa_A = \tilde{\kappa}_A + \kappa$ and $\kappa_B = \tilde{\kappa}_B + \kappa$. In analogy to reducing cost heterogeneity, an increase in general costs κ can then trigger the same qualitative trade-off between attenuating political polarization and improving candidate selection.

22. E.g., candidates may be forced by law to sell any shares of media stations that they own before running for office.

higher expected legacy. An increase in candidate B's or a decrease in candidate A's cost parameter then reduces the difference in manipulation intensities and, therefore, decreases polarization. Concerning the quality of candidate selection, the effect is reversed though. As discussed at the beginning of this section, naivety yields an advantage for the initially trailing candidate A, because naive voters excessively respond to newly arriving information. If costs are similar across candidates, A will, therefore, benefit from a higher likelihood to win the election compared to a situation in which all voters are sophisticated ($\tilde{\gamma}^* > \tilde{\gamma}^{SV}$). An increase in candidate B's cost parameter or a decrease in candidate A's cost parameter will reinforce A's advantage and impair candidate selection. Consequently changes in cost heterogeneity induce a trade-off between the quality of candidate selection and polarization.

Policy implications concerning the media access of candidates thus crucially depend on candidates' heterogeneity in costs. If candidates' costs to manipulate information differ substantially, institutions granting that no candidate has excessive power to influence or control media stations when campaigning are unambiguously welfare improving. If candidates resemble each other, however, the overall effect of a reduction in cost heterogeneity on welfare depends on the relative importance of preventing polarization and improving candidate selection for societal well-being.

1.7 Conclusions

In this paper, we have studied the welfare effects of information manipulation during democratic elections. We derive three main results. First, information manipulation will aggravate the polarization of political attitudes in a society. Second, information manipulation can improve candidate selection if the candidates' heterogeneity in terms of their costs to manipulate information is small. In this case, the welfare consequences depend on the relative importance of candidate selection and political polarization. If, however, candidates substantially differ in their costs, information manipulation will harm welfare in two ways – by impeding candidate selection and by aggravating political polarization. Third, we characterize conditions under which it is welfare improving to ensure an equal access to the media for both candidates.

A crucial assumption in our model is that some voters are naive when incorporating new information into their priors – they take public information at face value. While this assumption simplifies the analysis, it is also restrictive. However, our results do not rely on the exact specification of the bias, but rather on the idea that two phenomena are prevalent in an electorate. First, some voters trust information even if it is seemingly obvious that it is manipulated. Second, some voters overreact to new information compared to their prior. These characteristics are shared by several biases that have been discussed in the literature in various contexts. For example, individuals that receive new information from third parties are often described to be credulous (as for example in Baron, 1994; Grossman and Helpman, 1996; Ottaviani and Squintani, 2006; Kartik, Ottaviani, and Squintani, 2007; Inderst and Ottaviani, 2013; Little, 2017), overconfident (as in Ortoleva and Snowberg, 2015; Ogden, 2019), or to suffer from correlation or base-rate neglect (Enke and Zimmermann, 2017; Benjamin, Bodoh-Creed, and Rabin, 2019). Qualitatively our results should therefore transfer to settings in which a fraction of the electorate behaves in line with biases of these kinds.

Our results add to a lively debate about the impact of information manipulation in recent elections. In particular, they suggest that two recent phenomena – the increase of false information in political campaigns via social media and the polarization of political attitudes – are not independent. Instead, the polarization of political attitudes may have come about precisely because it became much less troublesome to spread incorrect information to large groups of individuals in an unfiltered way. For a better understanding of the reasons for political polarization, it should be of direct interest to test this assertion. Importantly, as argued by Allcott and Gentzkow (2017) and Lazer et al. (2018) there may also be a causal relation in the reversed direction – political polarization may be a perfect breeding ground for information manipulation. If both views are correct, information manipulation and political polarization might aggravate each other with detrimental consequences for social cohesion.

The interrelationship between polarization and the spreading of false information may also be influenced by further players in the political process. While

our model embraces the notion that political candidates at least indirectly spread false information, it may also originate from lobbies (Grossman and Helpman, 1996, 2001; Medina, 2019) or media stations (Gentzkow, Shapiro, and Stone, 2015). To what extent the information distortions caused by these different originators of biased information aggravate or attenuate each other is, however, hardly understood at all. Our signal-jamming approach with normally distributed beliefs to model information manipulation is tractable and, therefore, lends itself to add additional (strategic) players. This simplicity of the model could also be leveraged to study how various kinds of failures to account for the information structure affect voter behavior and polarization. For example, the model provides a natural framework to capture the idea that voters update selectively, i.e., they only update their beliefs about politicians if they receive news that confirms their political views.

Appendix 1.A Proofs

Proof of Proposition 1.1. The proof will proceed as follows: First, we derive the candidates' optimal manipulation intensities given that an interior equilibrium exists. Second, we show that there will exist an interior equilibrium in pure strategies if $\underline{c}$ and $\bar{m}$ are sufficiently large.

Suppose there exists an interior equilibrium. Then candidate B maximizes

$$\big(1-G_\gamma(\tilde{\gamma})\big)\cdot E\big[\,b+(1-\Sigma)\bar{q}_{B0}+\Sigma\cdot\big(q_B+\epsilon_B\big)\big|\,\gamma\,>\,\tilde{\gamma}\,\big]-c_B(m_B). \tag{1.A.1}$$

Define $X := q_B+\epsilon_B \sim N(\mu_X,\sigma^2)$ with $\mu_X := \bar{q}_{B0}$ and $\sigma^2 := \sigma_0^2+1$. Moreover, define $Y := q_A+\epsilon_A \sim N(\mu_Y,\sigma^2)$ with $\mu_Y := \bar{q}_{A0}$ and $\sigma^2 := \sigma_0^2+1$. Thus,

$$\gamma \;=\; X-Y \;\sim\; N(\mu_\gamma,\sigma_\gamma^2) \;=\; N\big(\mu_X-\mu_Y,2\sigma^2\big). \tag{1.A.2}$$

Let f_i denote the density and F_i the cdf of $i\in\{X,Y\}$. Then, we can compute the following conditional cdf:

$$\begin{aligned}
F_{X|X-Y>\tilde{\gamma}}(x) \;&:=\; \frac{P(X\,<\,x\cap X-Y\,>\,\tilde{\gamma})}{P(X-Y\,>\,\tilde{\gamma})}\\
&=\frac{1}{P(X-Y\,>\,\tilde{\gamma})}\int_{-\infty}^{x} f_X(u)\cdot F_Y(u-\tilde{\gamma})du\\
&=\frac{1}{1-G_\gamma(\tilde{\gamma})}\int_{-\infty}^{x} f_X(u)\cdot F_Y(u-\tilde{\gamma})du.
\end{aligned}$$

The corresponding density is given by

$$f_{X|X-Y>\tilde{\gamma}}(x) \;=\; \frac{1}{1-G_\gamma(\tilde{\gamma})}\cdot f_X(x)\cdot F_Y(x-\tilde{\gamma}),$$

so that

$$\begin{aligned}
E\big[\,\Sigma\big(q_B+\epsilon_B\big)\big|\,\gamma\,>\,\tilde{\gamma}\,\big] &= \Sigma E[X|X-Y\,>\,\tilde{\gamma}]\\
&=\frac{\Sigma}{1-G_\gamma(\tilde{\gamma})}\int_{-\infty}^{\infty} x\cdot f_X(x)\cdot F_Y(x-\tilde{\gamma})dx.
\end{aligned}$$

Therefore, candidate B's objective function (1.A.1) can be rewritten as

$$\big(1-G_\gamma(\tilde{\gamma})\big)[b+(1-\Sigma)\bar{q}_{B0}]+\Sigma\int_{-\infty}^{\infty} x\cdot f_X(x)\cdot F_Y(x-\tilde{\gamma})dx-c_B(m_B).$$

The first-order condition for m_B yields

$$-g_\gamma(\tilde{\gamma})\frac{\partial\tilde{\gamma}}{\partial m_B}[b+(1-\Sigma)\bar{q}_{B0}]-\frac{\partial\tilde{\gamma}}{\partial m_B}\Sigma\int_{-\infty}^{\infty}x\cdot f_X(x)\cdot f_Y(x-\tilde{\gamma})dx\;=\;c_B'(m_B). \tag{1.A.3}$$

The integral can be computed as follows:

$$\begin{aligned}
&\int_{-\infty}^{\infty}x\cdot f_X(x)\cdot f_Y(x-\tilde{\gamma})dx \qquad (1.A.4)\\
&=\int_{-\infty}^{\infty}\frac{x}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma^2}\right\}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-(\tilde{\gamma}+\mu_Y))^2}{2\sigma^2}\right\}dx\\
&=\int_{-\infty}^{\infty}\frac{x}{\sqrt{\pi\sigma^2}}\cdot g_\gamma(\tilde{\gamma})\cdot\exp\left\{-\frac{(2x-\tilde{\gamma}-\mu_X-\mu_Y)^2}{4\sigma^2}\right\}dx\\
&=g_\gamma(\tilde{\gamma})\cdot\int_{-\infty}^{\infty}x\cdot\frac{1}{\sqrt{\pi\sigma^2}}\exp\left\{-\frac{\left(x-\frac{\tilde{\gamma}+\mu_X+\mu_Y}{2}\right)^2}{\sigma^2}\right\}dx\\
&=g_\gamma(\tilde{\gamma})\cdot\frac{\tilde{\gamma}+\mu_X+\mu_Y}{2}.
\end{aligned}$$

Therefore, the first-order condition (1.A.3) can be rewritten as

$$\begin{aligned}
&-g_\gamma(\tilde{\gamma})\frac{\partial\tilde{\gamma}}{\partial m_B}\left(b+(1-\Sigma)\bar{q}_{B0}+\Sigma\frac{\mu_X+(\tilde{\gamma}+\mu_Y)}{2}\right)\;=\;c_B'(m_B)\\
\Leftrightarrow\qquad&-g_\gamma(\tilde{\gamma})\frac{\partial\tilde{\gamma}}{\partial m_B}\left(b+\bar{q}_{B0}+\frac{\tilde{\gamma}\Sigma}{2}+\frac{\Sigma(\bar{q}_{A0}-\bar{q}_{B0})}{2}\right)\;=\;c_B'(m_B),
\end{aligned}$$

which is identical to (1.9) as $\partial\tilde{\gamma}/\partial m_B=-1$ and $\mu_\gamma=\bar{q}_{B0}-\bar{q}_{A0}$.

Candidate A maximizes

$$G_\gamma(\tilde{\gamma})\cdot E\left[\,b+(1-\Sigma)\bar{q}_{A0}+\Sigma\cdot\left(q_A+\epsilon_A\right)\middle|\,\gamma\;<\;\tilde{\gamma}\right]-c_A(m_A). \tag{1.A.5}$$

We can construct the following conditional cdf:

$$\begin{aligned}
F_{Y|X-Y<\tilde{\gamma}}(y)\;&:=\frac{P(Y\;<\;y\cap X-Y\;<\;\tilde{\gamma})}{P(X-Y\;<\;\tilde{\gamma})}\\
&=\frac{1}{P(X-Y\;<\;\tilde{\gamma})}\int_{-\infty}^{y}f_Y(v)\cdot F_X(v+\tilde{\gamma})dv
\end{aligned}$$

$$= \frac{1}{G_\gamma(\tilde{\gamma})} \int_{-\infty}^{y} f_Y(v) \cdot F_X(v+\tilde{\gamma}) dv,$$

which has the density

$$f_{Y|X-Y<\tilde{\gamma}}(y) = \frac{1}{G_\gamma(\tilde{\gamma})} f_Y(y) \cdot F_X(y+\tilde{\gamma}).$$

Hence,

$$E[Y|X-Y < \tilde{\gamma}] = \frac{1}{G_\gamma(\tilde{\gamma})} \int_{-\infty}^{\infty} y \cdot f_Y(y) \cdot F_X(y+\tilde{\gamma}) dy,$$

so that candidate A's objective function (1.A.5) can be rewritten as

$$G_\gamma(\tilde{\gamma}) \cdot [b + (1-\Sigma)\bar{q}_{A0}] + \Sigma \cdot \int_{-\infty}^{\infty} y \cdot f_Y(y) \cdot F_X(y+\tilde{\gamma}) dy - c_A(m_A).$$

As first-order condition for m_A, we obtain

$$g_\gamma(\tilde{\gamma}) \frac{\partial \tilde{\gamma}}{\partial m_A} \cdot [b + (1-\Sigma)\bar{q}_{A0}] + \Sigma \frac{\partial \tilde{\gamma}}{\partial m_A} \cdot \int_{-\infty}^{\infty} y \cdot f_Y(y) \cdot f_X(y+\tilde{\gamma}) dy = c'_A(m_A). \tag{1.A.6}$$

In analogy to (1.A.4), the integral can be computed as

$$\int_{-\infty}^{\infty} y \cdot f_Y(y) \cdot f_X(y+\tilde{\gamma}) dy = g_\gamma(\tilde{\gamma}) \cdot \frac{-\tilde{\gamma} + \mu_X + \mu_Y}{2}$$

so that the first-order condition (1.A.6) can be rewritten as

$$\begin{aligned} & g_\gamma(\tilde{\gamma}) \frac{\partial \tilde{\gamma}}{\partial m_A} \left(b + (1-\Sigma)\bar{q}_{A0} + \Sigma \frac{-\tilde{\gamma} + \mu_X + \mu_Y}{2} \right) = c'_A(m_A) \\ \Leftrightarrow \quad & g_\gamma(\tilde{\gamma}) \frac{\partial \tilde{\gamma}}{\partial m_A} \left(b + \mu_Y - \frac{\Sigma}{2} \left(\tilde{\gamma} - \mu_\gamma \right) \right) = c'_A(m_A), \end{aligned}$$

which is identical to (1.8) as $\partial \tilde{\gamma} / \partial m_A = 1$. The equilibrium cutoff (1.10) is obtained from equation (1.7) together with the fact that sophisticated voters correctly form beliefs about the two candidates' equilibrium manipulation intensities, i.e., $\hat{m}_j = m_j^*$ $(j = A, B)$.

To prove existence of an interior equilibrium in pure strategies, we proceed in two steps. In step 1, we prove that there exists a pure-strategy equilibrium if $\underline{c}$ is

sufficiently large. In step 2, we show that this equilibrium will be interior if $\bar{m}$ is sufficiently large.

Step 1: To prove existence, we consider a two-stage auxiliary game. At the second stage, nature determines the realization of the random variable γ. At the first stage, three players, $\hat{A}$, $\hat{B}$ and $\hat{S}$, simultaneously choose their actions. The objective functions of players $\hat{A}$ and $\hat{B}$ are given by (1.A.1) and (1.A.5), and their action spaces by $[0, \bar{m}]$. Hence, they have the same objective functions and the same action spaces as the respective candidates in the original game. Player $\hat{S}$ chooses $\hat{m}_A, \hat{m}_B \in [0, \bar{m}]$ and has the objective function[23]

$$\max_{\hat{m}_A, \hat{m}_B} -\left(m_A - \hat{m}_A\right)^2 - \left(m_B - \hat{m}_B\right)^2. \tag{1.A.7}$$

We first argue that any strategy profile (m_A, m_B) that is part of a Nash equilibrium in the auxiliary game also constitutes an equilibrium in the original game. By the definition of the equilibrium, the following conditions have to hold:

(1) Both candidates play mutually best responses given that voters vote sincerely.

(2) Naive voters believe that candidates do not invest in manipulation of beliefs whereas sophisticated voters correctly anticipate manipulation intensities along the equilibrium path.

Equation (1.A.7) implies that the best response of $\hat{S}$ is given by (m_A, m_B) such that $\hat{m}_j^* = m_j^*$ holds for $j \in \{A, B\}$ in any Nash equilibrium of the auxiliary game. Moreover, the payoffs in the auxiliary game are identical to the ones that arise in the original game if voters vote sincerely, sophisticated voters hold correct beliefs about manipulation intensities, and naive voters believe manipulation intensities to be zero. Consequently any strategy profile (m_A, m_B) that is part of a Nash equilibrium in the auxiliary game constitutes an equilibrium in the original game.

We now provide sufficient conditions for the existence of a pure-strategy Nash equilibrium in the auxiliary game. The objective function of player $\hat{S}$ is concave in his

23. For a similar argument see Dato, Grunewald, Müller, and Strack (2017) and Eliaz and Spiegler (2020).

own actions and continuous in his own and the other players' actions. Furthermore, his strategy space is a compact and convex subset of an Euclidean space. Additionally, consider the second derivative of the candidates' objective functions in the original game:

$$-g'_\gamma(\tilde{\gamma})\left(b+\bar{q}_{B0}+\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)-\frac{\Sigma}{2}g_\gamma(\tilde{\gamma})-\kappa_B c''(m_B)$$

and

$$g'_\gamma(\tilde{\gamma})\left(b+\bar{q}_{A0}-\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)-\frac{\Sigma}{2}g_\gamma(\tilde{\gamma})-\kappa_A c''(m_A).$$

Thus, if[24]

$$\sup_x\left\{|g'_\gamma(x)|(b+\bar{q}_{j0}+\frac{\Sigma}{2}|x-\mu_\gamma|)-\frac{\Sigma}{2}g_\gamma(x)\right\}\ <\ \kappa_j\cdot\underline{c} \tag{1.A.8}$$

for $j=A,B$, then the candidates' objective function is concave such that we can apply the Debreu-Glicksberg-Fan Existence Theorem (Debreu, 1952; Fan, 1952; Glicksberg, 1952) for infinite games to obtain the existence of a pure-strategy Nash equilibrium in the auxiliary game. The subsequent results are based on condition (1.A.8) being satisfied, implying concave objective functions of both candidates.

Step 2: We show that the equilibrium is interior if $\bar{m}$ is sufficiently large. According to (1.8) and (1.9), the first derivatives of the candidates' objective functions in equilibrium are

$$g_\gamma\left(\tilde{\gamma}^*\right)\left[b+\bar{q}_{A0}-\frac{\Sigma}{2}\left(\tilde{\gamma}^*-\mu_\gamma\right)\right]-\kappa_A c'\left(m_A^*\right)$$

and

$$g_\gamma\left(\tilde{\gamma}^*\right)\left[b+\bar{q}_{B0}+\frac{\Sigma}{2}\left(\tilde{\gamma}^*-\mu_\gamma\right)\right]-\kappa_B c'\left(m_B^*\right).$$

Consider $m_A^*=m_B^*=0$, first. In this case, it has to hold that

$$g_\gamma(\tilde{\gamma}^*)\left(b+\bar{q}_{A0}-\frac{\Sigma}{2}\left(\tilde{\gamma}^*-\mu_\gamma\right)\right)-c'_A(0)\ \leq\ 0$$

24. Similar to the standard assumption in Lazear-Rosen type tournaments, the players' cost functions have to be sufficiently convex to guarantee the existence of an interior solution; see, e.g., Lazear and Rosen (1981), Nalebuff and Stiglitz (1983), and Schöttner (2008).

and

$$g_\gamma(\tilde{\gamma}^*)\left(b+\bar{q}_{B0}+\frac{\Sigma}{2}\left(\tilde{\gamma}^*-\mu_\gamma\right)\right)-c_B'(0)\ \leq\ 0.$$

As, by assumption, $c_A'(0)=c_B'(0)=0$, this yields

$$b+\bar{q}_{A0}-\frac{\Sigma}{2}\left(\tilde{\gamma}^*-\mu_\gamma\right)\ \leq\ 0$$

and

$$b+\bar{q}_{B0}+\frac{\Sigma}{2}\left(\tilde{\gamma}^*-\mu_\gamma\right)\ \leq\ 0.$$

Adding up both sides of the inequalities yields

$$2b+\bar{q}_{A0}+\bar{q}_{B0}\ \leq\ 0,$$

which is never fulfilled.

For the other cases, it will be helpful to rewrite the first derivatives of the objective functions in equilibrium to

$$g_\gamma\left(\tilde{\gamma}^*\right)\left[b+\frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}-\frac{\Sigma}{2}\left(\tilde{\gamma}^*+\frac{1-\Sigma}{\Sigma}\mu_\gamma\right)\right]-\kappa_A c'\left(m_A^*\right)$$

and

$$g_\gamma\left(\tilde{\gamma}^*\right)\left[b+\frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}+\frac{\Sigma}{2}\left(\tilde{\gamma}^*+\frac{1-\Sigma}{\Sigma}\mu_\gamma\right)\right]-\kappa_B c'\left(m_B^*\right).$$

Consider $m_A^*>0$ and $m_B^*=0$, now. Then it has to hold that

$$\begin{aligned}
&g_\gamma\left(\tilde{\gamma}^*\right)\left[b+\frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}+\frac{\Sigma}{2}\left(\tilde{\gamma}^*+\frac{1-\Sigma}{\Sigma}\mu_\gamma\right)\right]-\kappa_B c'(0)\ \leq\ 0\\
\Rightarrow\ &\frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}+\frac{\Sigma}{2}\left(\tilde{\gamma}^*+\frac{1-\Sigma}{\Sigma}\mu_\gamma\right)\ \leq\ 0\\
\Leftrightarrow\ &\tilde{\gamma}^*+\frac{1-\Sigma}{\Sigma}\mu_\gamma\ \leq\ -\frac{\bar{q}_{A0}+\bar{q}_{B0}}{\Sigma},
\end{aligned}$$

because otherwise there exists a profitable deviation for candidate B. Because $-\frac{\bar{q}_{A0}+\bar{q}_{B0}}{\Sigma}<0$ we obtain

$$\tilde{\gamma}^*\ <\ -\frac{1-\Sigma}{\Sigma}\mu_\gamma\ <\ 0,$$

which implies

$$(1-\Sigma)\,\mu_\gamma+\Sigma\tilde{\gamma}^*\ <\ 0.$$

Since $G_\theta(\cdot)$ is a strictly increasing function we can deduce that

$$G_\theta\left((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^*\right) < G_\theta(0) = \frac{1}{2},$$

where the last equality stems from the assumption, that g_θ is symmetric with a unique mode at zero. As (1.7) holds in any equilibrium, this implies that

$$G_\theta\left(\tilde{\gamma}^* - m_A^*\right) > \frac{1}{2} \iff \tilde{\gamma}^* - m_A^* > 0 \Rightarrow \tilde{\gamma}^* > 0$$

because $m_A^* > 0$, which yields a contradiction.

Consider the case $m_B^* > 0$ and $m_A^* = 0$, now. In this case, it has to hold that

$$\begin{aligned} & g_\gamma\left(\tilde{\gamma}^*\right)\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2} - \frac{\Sigma}{2}\left(\tilde{\gamma}^* + \frac{1-\Sigma}{\Sigma}\mu_\gamma\right)\right] - \kappa_A c'(0) \leq 0 \\ \iff & b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2} \leq \frac{\Sigma}{2}\left(\tilde{\gamma}^* + \frac{1-\Sigma}{\Sigma}\mu_\gamma\right) \\ \iff & b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2} - \frac{(1-\Sigma)\mu_\gamma}{2} \leq \frac{\Sigma}{2}\tilde{\gamma}^* \end{aligned}$$

because otherwise candidate A has a profitable deviation. As $\bar{q}_{A0} + \bar{q}_{B0} > (1-\Sigma)\mu_\gamma$, we can conclude that $\tilde{\gamma}^* \geq 0$.

This implies

$$\begin{aligned} & (1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^* > 0 \\ \iff & G_\theta\left((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^*\right) > \frac{1}{2} = G_\theta(0), \end{aligned}$$

where the last equality stems from the assumption, that g_θ is symmetric with a unique mode at zero. As (1.7) holds in every equilibrium, this implies that

$$G_\theta\left(\tilde{\gamma}^* + m_B^*\right) < \frac{1}{2} \iff \tilde{\gamma}^* + m_B^* < 0 \iff \tilde{\gamma}^* < -m_B^* < 0,$$

a contradiction.

Note that we have shown that either manipulation intensities are interior or at their upper bound. Consequently, we can conclude that

$$g_\gamma\left(\tilde{\gamma}^*\right)\left[b + \bar{q}_{A0} - \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right)\right] \geq \kappa_A c'\left(m_A^*\right)$$

and

$$g_\gamma\left(\tilde{\gamma}^*\right)\left[b + \bar{q}_{B0} + \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right)\right] \geq \kappa_B c'\left(m_B^*\right).$$

Summing up the two sides of the inequalities yields

$$g_\gamma\left(\tilde{\gamma}^*\right)\left(2b+\bar{q}_{A0}+\bar{q}_{B0}\right) \geq \kappa_A c'\left(m_A^*\right)+\kappa_B c'\left(m_B^*\right).$$

Since g_γ is the density of a normally distributed random variable with mean μ_γ, it attains its maximum at μ_γ. Consequently,

$$g_\gamma\left(\mu_\gamma\right)\left(2b+\bar{q}_{A0}+\bar{q}_{B0}\right) \geq \kappa_A c'\left(m_A^*\right)+\kappa_B c'\left(m_B^*\right)$$

because

$$g_\gamma\left(\mu_\gamma\right)\left(2b+\bar{q}_{A0}+\bar{q}_{B0}\right) \geq g_\gamma\left(\tilde{\gamma}^*\right)\left(2b+\bar{q}_{A0}+\bar{q}_{B0}\right).$$

As $\lim_{m\to\infty} c'(m)=\infty$ $(j=A,B)$ and $g_\gamma\left(\mu_\gamma\right)\left(2b+\bar{q}_{A0}+\bar{q}_{B0}\right)$ is finite, we can always find a sufficiently large but finite $\bar{m}$ such that

$$g_\gamma\left(\mu_\gamma\right)\left(2b+\bar{q}_{A0}+\bar{q}_{B0}\right) < \kappa_A c'(\bar{m})$$

and

$$g_\gamma\left(\mu_\gamma\right)\left(2b+\bar{q}_{A0}+\bar{q}_{B0}\right) < \kappa_B c'(\bar{m}),$$

which implies that the equilibrium will be interior if $\bar{m}$ is sufficiently large. □

Proof of Proposition 1.2. It remains to derive the equilibrium expression for political polarization, $\mathscr{P}\left(m_A^*, m_B^*\right)$. The mean of Δu is given by

$$\begin{aligned}\mu_{\Delta u} &= \int_{-\infty}^{\infty} x h\left(x|\gamma\right)dx = \int_{-\infty}^{\infty} x(\alpha h^N\left(x|\gamma\right)+(1-\alpha)h^S\left(x|\gamma\right))dx \\ &= \alpha(\bar{q}_{A1}^N-\bar{q}_{B1}^N)+(1-\alpha)(\bar{q}_{A1}^S-\bar{q}_{B1}^S),\end{aligned}$$

where the last equality follows from the fact that the density $h^i\left(\cdot|\gamma\right)$ has mean $\bar{q}_{A1}^i-\bar{q}_{B1}^i$. The variance of Δu can be computed as follows:

$$\begin{aligned}\sigma_{\Delta u}^2 &= \int_{-\infty}^{\infty} x^2 h(x|\gamma)dx-\mu_{\Delta u}^2 \\ &= \alpha\int_{-\infty}^{\infty} x^2 h^N(x|\gamma)dx+(1-\alpha)\int_{-\infty}^{\infty} x^2 h^S(x|\gamma)dx-\mu_{\Delta u}^2 \\ &= \alpha\left[(\bar{q}_{A1}^N-\bar{q}_{B1}^N)^2+\sigma_\theta^2\right]+(1-\alpha)\left[(\bar{q}_{A1}^S-\bar{q}_{B1}^S)^2+\sigma_\theta^2\right]-\mu_{\Delta u}^2\end{aligned}$$

$$= \sigma_\theta^2 + \alpha(1-\alpha)\left[(\bar{q}_{A1}^N - \bar{q}_{B1}^N) - (\bar{q}_{A1}^S - \bar{q}_{B1}^S)\right]^2.$$

As $\bar{q}_{A1}^N - \bar{q}_{B1}^N = m_A^* - m_B^* - \gamma$ and $\bar{q}_{A1}^S - \bar{q}_{B1}^S = -(1-\Sigma)\mu_\gamma - \Sigma\gamma$, we obtain

$$\sigma_{\Delta u}^2 = \sigma_\theta^2 + \alpha(1-\alpha)\,(m_A^* - m_B^* + (1-\Sigma)\left(\mu_\gamma - \gamma\right))^2, \tag{1.A.9}$$

Taking the expectation with respect to γ yields

$$\begin{aligned}
\mathbb{E}_\gamma\left[\sigma_{\Delta u}^2\right] &= \mathbb{E}_\gamma\left[\sigma_\theta^2 + \alpha(1-\alpha)\,(m_A^* - m_B^* + (1-\Sigma)\left(\mu_\gamma - \gamma\right))^2\right] \\
&= \mathbb{E}_\gamma\left[\sigma_\theta^2 + \alpha(1-\alpha)\left[(m_A^* - m_B^*)^2 + (1-\Sigma)^2\left(\mu_\gamma - \gamma\right)^2\right]\right] \\
&= \sigma_\theta^2 + \alpha(1-\alpha)\left[(m_A^* - m_B^*)^2 + (1-\Sigma)^2\,\sigma_\gamma^2\right].
\end{aligned}$$

□

Proof of Lemma 1.1. We first introduce another lemma that is useful for proving Lemma 1.1 and the subsequent propositions:

Lemma 1.2. *Consider $\tilde{\gamma}'$ and $\tilde{\gamma}''$. If $\tilde{\gamma}^{SV} < \tilde{\gamma}' < \tilde{\gamma}''$ or $\tilde{\gamma}'' < \tilde{\gamma}' < \tilde{\gamma}^{SV}$, then $\mathscr{E}\left(\tilde{\gamma}'\right) > \mathscr{E}\left(\tilde{\gamma}''\right)$.*

Proof of Lemma 1.2. According to Definition (1.13), $\mathscr{E}\left(\tilde{\gamma}\right)$ is given by

$$E[q_A|\gamma < \tilde{\gamma}] \cdot G_\gamma\left(\tilde{\gamma}\right) + E[q_B|\gamma \geq \tilde{\gamma}] \cdot \left(1 - G_\gamma\left(\tilde{\gamma}\right)\right).$$

To prove the lemma suppose to the contrary that $\mathscr{E}\left(\tilde{\gamma}'\right) \leq \mathscr{E}\left(\tilde{\gamma}''\right)$. Therefore,

$$\begin{aligned}
&E\left[q_A|\gamma < \tilde{\gamma}'\right] \cdot G_\gamma\left(\tilde{\gamma}'\right) + E\left[q_B|\tilde{\gamma}' \leq \gamma\right] \cdot \left(1 - G_\gamma\left(\tilde{\gamma}'\right)\right) \\
\leq\, &E\left[q_A|\gamma < \tilde{\gamma}''\right] \cdot G_\gamma\left(\tilde{\gamma}''\right) + E\left[q_B|\tilde{\gamma}'' \leq \gamma\right] \cdot \left(1 - G_\gamma\left(\tilde{\gamma}''\right)\right).
\end{aligned}$$

This inequality can be rewritten as

$$\begin{aligned}
&\int_{-\infty}^{\tilde{\gamma}'} E[q_A|\gamma]\, g_\gamma\left(\gamma\right) \mathrm{d}\gamma + \int_{\tilde{\gamma}'}^{\infty} E[q_B|\gamma]\, g_\gamma\left(\gamma\right) \mathrm{d}\gamma \\
\leq\, &\int_{-\infty}^{\tilde{\gamma}''} E[q_A|\gamma]\, g_\gamma\left(\gamma\right) \mathrm{d}\gamma + \int_{\tilde{\gamma}''}^{\infty} E[q_B|\gamma]\, g_\gamma\left(\gamma\right) \mathrm{d}\gamma
\end{aligned}$$

$$\Leftrightarrow \quad \int_{\tilde{\gamma}'}^{\tilde{\gamma}''} E[q_B - q_A|\gamma] g_\gamma(\gamma) \mathrm{d}\gamma \leq 0. \tag{1.A.10}$$

As (e.g., by DeGroot (2005))

$$E[q_B - q_A|\gamma] = E[q_B - q_A] + \frac{Var[q_B - q_A]}{Var[\gamma]} \cdot (\gamma - E[\gamma]) \tag{1.A.11}$$

with $E[q_B - q_A] = E[\gamma] = \mu_\gamma$ and $Var[q_B - q_A] = 2\sigma_0^2$ and $Var[\gamma]$ being displayed in (1.A.2) so that $Var[q_B - q_A]/Var[\gamma] = \Sigma$, inequality (1.A.10) can be rewritten as

$$\int_{\tilde{\gamma}'}^{\tilde{\gamma}''} \left((1-\Sigma)\mu_\gamma + \Sigma\gamma\right) g_\gamma(\gamma) \mathrm{d}\gamma \leq 0.$$

If $\tilde{\gamma}^{SV} < \tilde{\gamma}' < \tilde{\gamma}''$, this inequality can only be fulfilled if $(1-\Sigma)\mu_\gamma + \Sigma\gamma$ adopts negative values for some γ such that $\tilde{\gamma}' \leq \gamma \leq \tilde{\gamma}''$. This is the case if and only if

$$\gamma < -\frac{1-\Sigma}{\Sigma}\mu_\gamma = \tilde{\gamma}^{SV} < \tilde{\gamma}',$$

which is never fulfilled.

If on the other hand $\tilde{\gamma}'' < \tilde{\gamma}' < \tilde{\gamma}^{SV}$, this inequality can only be fulfilled if $(1-\Sigma)\mu_\gamma + \Sigma\gamma$ adopts positive values for some γ such that $\tilde{\gamma}'' \leq \gamma \leq \tilde{\gamma}'$. This is the case if and only if

$$\gamma > -\frac{1-\Sigma}{\Sigma}\mu_\gamma = \tilde{\gamma}^{SV} > \tilde{\gamma}',$$

which is also never fulfilled. □

Now, we can prove Lemma 1.1. Suppose, to the contrary, $\tilde{\gamma}^{NM} \leq \tilde{\gamma}^{SV} < 0$. This implies

$$(1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^{NM} \leq (1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^{SV} = 0,$$

where the last equality follows from (1.18). However, by (1.15),

$$\tilde{\gamma}^{NM} \geq 0,$$

which yields a contradiction. Hence, $\tilde{\gamma}^{NM} > \tilde{\gamma}^{SV}$. The claim $\mathscr{E}(\tilde{\gamma}^{NM}) < \mathscr{E}(\tilde{\gamma}^{SV})$ then directly follows from $\tilde{\gamma}^{NM} \neq \tilde{\gamma}^{SV}$.

Implicit differentiation of (1.15) yields

$$\frac{\partial \tilde{\gamma}^{NM}}{\partial \alpha} = \frac{G_\theta\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM}\right) - G_\theta\left(\tilde{\gamma}^{NM}\right)}{(1-\alpha)\,g_\theta\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM}\right)\cdot\Sigma + \alpha g_\theta(\tilde{\gamma}^{NM})}.$$

Thus, $\frac{\partial \tilde{\gamma}^{NM}}{\partial \alpha} > 0$ if

$$G_\theta\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM}\right) > G_\theta\left(\tilde{\gamma}^{NM}\right).$$

$\tilde{\gamma}^{NM} > \tilde{\gamma}^{SV}$ together with equation (1.18) implies that $G_\theta\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM}\right) > \frac{1}{2}$. By (1.15) this implies $G_\theta\left(\tilde{\gamma}^{NM}\right) < \frac{1}{2}$ so that $\frac{\partial \tilde{\gamma}^{NM}}{\partial \alpha} > 0$ holds. Thus, according to Lemma 1.2, $\partial \mathcal{E}(\tilde{\gamma}^{NM})/\partial\alpha < 0$. □

Proof of Proposition 1.3. First, we state two lemmas that are useful for proving this Proposition.

Lemma 1.3. *Let $\kappa_A = \kappa_B$. Then $m_B^* \gtreqless m_A^*$ if and only if $\tilde{\gamma}^* \gtreqless \tilde{\gamma}^{SV}$.*

Proof. Given $\kappa_B = \kappa_A$, Proposition 1.1 shows that $m_B^* \gtreqless m_A^*$ if and only if

$$\begin{aligned} & b + \bar{q}_{B0} + \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right) \gtreqless b + \bar{q}_{A0} - \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right) \\ \Leftrightarrow\ & \bar{q}_{B0} + \Sigma\left(\tilde{\gamma}^* - \mu_\gamma\right) \gtreqless \bar{q}_{A0} \\ \Leftrightarrow\ & \tilde{\gamma}^* \gtreqless -\frac{1-\Sigma}{\Sigma}\mu_\gamma = \tilde{\gamma}^{SV}. \end{aligned}$$

□

Lemma 1.4. *$m_B^* \gtreqless m_A^*$ if and only if $\tilde{\gamma}^* \lesseqgtr \tilde{\gamma}^{NM}$.*

Proof. We prove the lemma by proving the following statement, as it immediately implies the other direction of the lemma's claim: If $m_B^* \gtreqless m_A^*$, then $\tilde{\gamma}^* \lesseqgtr \tilde{\gamma}^{NM}$.

Consider $m_B^* > m_A^*$ and suppose, to the contrary, $\tilde{\gamma}^* \geq \tilde{\gamma}^{NM}$, which implies

$$(1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^* \geq (1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM},$$

and, due to (1.10) and (1.15),

$$\tilde{\gamma}^{NM} \geq \tilde{\gamma}^* - \left(m_A^* - m_B^*\right).$$

As, by assumption, $\tilde{\gamma}^* \geq \tilde{\gamma}^{NM}$, we must have that $m_A^* \geq m_B^*$, which contradicts the assumption that $m_B^* > m_A^*$.

In strict analogy, we can show that $m_B^* < m_A^*$ implies $\tilde{\gamma}^* > \tilde{\gamma}^{NM}$, as – due to (1.10) and (1.15) – $\tilde{\gamma}^* < \tilde{\gamma}^{NM}$ would contradict the assumption $m_B^* < m_A^*$. For the remaining case of $m_B^* = m_A^*$, $\tilde{\gamma}^* = \tilde{\gamma}^{NM}$ follows directly from the cutoff conditions. □

We now proceed to the proof of Proposition 1.3. We first show that, if $\kappa_A = \kappa_B$, then $\tilde{\gamma}^* > \tilde{\gamma}^{SV}$. Suppose, to the contrary, $\tilde{\gamma}^* \leq \tilde{\gamma}^{SV}$. By Lemma 1.1 this implies $\tilde{\gamma}^* < \tilde{\gamma}^{NM}$. Therefore,

$$(1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^* \;<\; (1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM}.$$

By the cutoff conditions (1.10) and (1.15) this implies

$$\tilde{\gamma}^* + \left(m_B^* - m_A^*\right) \;>\; \tilde{\gamma}^{NM}.$$

By Lemma 1.3, $\tilde{\gamma}^* \leq \tilde{\gamma}^{SV}$ implies $m_B^* \leq m_A^*$. Consequently

$$\tilde{\gamma}^* \;>\; \tilde{\gamma}^{NM},$$

which contradicts $\tilde{\gamma}^* < \tilde{\gamma}^{NM}$. Hence, $\tilde{\gamma}^* > \tilde{\gamma}^{SV}$ must hold. Due to Lemma 1.3, we get that $m_B^* > m_A^*$, which implies by Lemma 1.4 that $\tilde{\gamma}^* < \tilde{\gamma}^{NM}$. To sum up, we have $\tilde{\gamma}^{SV} < \tilde{\gamma}^* < \tilde{\gamma}^{NM}$ and, therefore, $\mathscr{E}(\tilde{\gamma}^*) > \mathscr{E}(\tilde{\gamma}^{NM})$ according to Lemma 1.2. However, $m_B^* \neq m_A^*$ implies $\mathscr{P}(m_A^*, m_B^*) > \mathscr{P}^{NM}$. Hence, there exists a critical value $\beta = \beta^*$ such that $\mathscr{W}^* = \mathscr{W}^{NM}$, and $\mathscr{W}^* \gtrless \mathscr{W}^{NM}$ if $\beta \lessgtr \beta^*$. □

Proof of Proposition 1.4. We start by proving two lemmas.

Lemma 1.5. *If $\kappa_A > \kappa_B$, then $m_B^* > m_A^*$.*

Proof. Suppose to the contrary $m_A^* \geq m_B^*$. Subtracting (1.9) from (1.8) yields

$$-g_\gamma\left(\tilde{\gamma}^*\right)\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^*\right) \;=\; \kappa_A c'\left(m_A^*\right) - \kappa_B c'\left(m_B^*\right).$$

From our assumptions we conclude $\kappa_A c'\left(m_A^*\right) - \kappa_B c'\left(m_B^*\right) > 0$ and, thus,

$$(1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^* \;<\; 0.$$

We obtain that $\tilde{\gamma}^* < \tilde{\gamma}^{SV}$, because $(1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{SV} = 0$. The cutoff condition (1.10) yields that

$$\tilde{\gamma}^* - \left(m_A^* - m_B^*\right) > 0 \iff m_A^* - m_B^* < \tilde{\gamma}^*,$$

which is a contradiction as $\tilde{\gamma}^* < \tilde{\gamma}^{SV} = -\frac{1-\Sigma}{\Sigma}\mu_\gamma < 0$. □

Lemma 1.6. $\tilde{\gamma}^{NM} < 0$.

Proof. By Lemma 1.1 we can infer that

$$(1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM} > (1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{SV} = 0$$

By (1.15) we can conclude that $\tilde{\gamma}^{NM} < 0$. □

We now turn to the proof of the claim in the Proposition. Lemmas 1.1, 1.4 and 1.5 imply that we can focus on the following equilibrium cutoff constellations

(a) $\tilde{\gamma}^{SV} < \tilde{\gamma}^* < \tilde{\gamma}^{NM}$ and (b) $\tilde{\gamma}^* < \tilde{\gamma}^{SV} < \tilde{\gamma}^{NM}$.

Lemma 1.2 directly implies the claim for constellation (a). Hence, consider constellation (b). Suppose $\mathscr{E}\left(\tilde{\gamma}^*\right) \leq \mathscr{E}\left(\tilde{\gamma}^{NM}\right)$. We infer from the steps in the proof of Lemma 1.2 that

$$\int_{\tilde{\gamma}^*}^{\tilde{\gamma}^{NM}} E[q_B - q_A|\gamma]\, g_\gamma(\gamma)\,\mathrm{d}\gamma \leq 0$$

$$\iff \int_{\tilde{\gamma}^{SV}}^{\tilde{\gamma}^{NM}} E[q_B - q_A|\gamma]\, g_\gamma(\gamma)\,\mathrm{d}\gamma \leq -\int_{\tilde{\gamma}^*}^{\tilde{\gamma}^{SV}} E[q_B - q_A|\gamma]\, g_\gamma(\gamma)\,\mathrm{d}\gamma \leq 0.$$

From the proof of Lemma 1.2 we know that $E[q_B - q_A|\gamma] \geq 0$ for all γ such that $\tilde{\gamma}^{SV} \leq \gamma \leq \tilde{\gamma}^{NM}$, and $E[q_B - q_A|\gamma] \leq 0$ for all γ such that $\tilde{\gamma}^* \leq \gamma \leq \tilde{\gamma}^{SV}$. Additionally, Lemma 1.6 implies that $\tilde{\gamma}^{NM} < \mu_\gamma$, and γ is normally distributed with mean μ_γ. Therefore, $g_\gamma(\gamma)$ is strictly increasing on the interval $\left[\tilde{\gamma}^*, \tilde{\gamma}^{NM}\right]$ and, hence,

$$\int_{\tilde{\gamma}^{SV}}^{\tilde{\gamma}^{NM}} E[q_B - q_A|\gamma]\, g_\gamma\left(\tilde{\gamma}^{SV}\right)\mathrm{d}\gamma < \int_{\tilde{\gamma}^{SV}}^{\tilde{\gamma}^{NM}} E[q_B - q_A|\gamma]\, g_\gamma(\gamma)\,\mathrm{d}\gamma$$

and

$$-\int_{\tilde{\gamma}^*}^{\tilde{\gamma}^{SV}} E[q_B - q_A|\gamma] g_\gamma(\gamma)\,\mathrm{d}\gamma \;<\; -\int_{\tilde{\gamma}^*}^{\tilde{\gamma}^{SV}} E[q_B - q_A|\gamma] g_\gamma\left(\tilde{\gamma}^{SV}\right)\mathrm{d}\gamma.$$

Consequently, if $\mathscr{E}\left(\tilde{\gamma}^*\right) \leq \mathscr{E}\left(\tilde{\gamma}^{NM}\right)$, then

$$\begin{aligned}
&\int_{\tilde{\gamma}^{SV}}^{\tilde{\gamma}^{NM}} E[q_B - q_A|\gamma] g_\gamma\left(\tilde{\gamma}^{SV}\right)\mathrm{d}\gamma \;<\; -\int_{\tilde{\gamma}^*}^{\tilde{\gamma}^{SV}} E[q_B - q_A|\gamma] g_\gamma\left(\tilde{\gamma}^{SV}\right)\mathrm{d}\gamma \\
\Leftrightarrow &\int_{\tilde{\gamma}^*}^{\tilde{\gamma}^{NM}} E[q_B - q_A|\gamma]\,\mathrm{d}\gamma \;<\; 0 \\
\Leftrightarrow &\int_{\tilde{\gamma}^*}^{\tilde{\gamma}^{NM}} ((1-\Sigma)\,\mu_\gamma + \Sigma\gamma)\mathrm{d}\gamma \;<\; 0,
\end{aligned}$$

where the last step follows from the steps in the proof of Lemma 1.2. Solving the integral leads to

$$\begin{aligned}
&\left(\tilde{\gamma}^{NM} - \tilde{\gamma}^*\right)\cdot(1-\Sigma)\,\mu_\gamma + \frac{\Sigma}{2}\left(\tilde{\gamma}^{NM} - \tilde{\gamma}^*\right)\left(\tilde{\gamma}^{NM} + \tilde{\gamma}^*\right) \;<\; 0 \\
\Leftrightarrow\; &\tilde{\gamma}^{NM} - \tilde{\gamma}^{SV} \;<\; \tilde{\gamma}^{SV} - \tilde{\gamma}^*.
\end{aligned}$$

A sufficient condition for $\mathscr{E}\left(\tilde{\gamma}^*\right) > \mathscr{E}\left(\tilde{\gamma}^{NM}\right)$ is, hence, given by

$$\tilde{\gamma}^{SV} - \tilde{\gamma}^* \;\leq\; \tilde{\gamma}^{NM} - \tilde{\gamma}^{SV}.$$

The rest of the proof will provide a sufficient condition for this inequality to hold. We start with the case of $\alpha = \frac{1}{2}$, where we can solve for $\tilde{\gamma}^{NM} - \tilde{\gamma}^{SV}$ explicitly. Then, we derive an upper bound for $\tilde{\gamma}^{SV} - \tilde{\gamma}^*$. Next, we provide conditions under which $\tilde{\gamma}^{NM} - \tilde{\gamma}^{SV}$ exceeds this upper bound. In a final step, we will show that $\tilde{\gamma}^{NM} - \tilde{\gamma}^{SV}$ increases in α, but the upper bound of $\tilde{\gamma}^{SV} - \tilde{\gamma}^*$ is independent of α.

If $\alpha = \frac{1}{2}$, then (1.15) will reduce to

$$\begin{aligned}
\frac{1}{2} \;&=\; \frac{1}{2}\cdot G_\theta\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM}\right) + \frac{1}{2}\cdot G_\theta\left(\tilde{\gamma}^{NM}\right) \\
\Leftrightarrow G_\theta\left(\tilde{\gamma}^{NM}\right) \;&=\; 1 - G_\theta\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{NM}\right).
\end{aligned}$$

Because G_θ is symmetric around 0 and g_θ has full support, we can infer

$$\begin{aligned} & G_\theta\left(\tilde{\gamma}^{NM}\right) = G_\theta\left(-(1-\Sigma)\,\mu_\gamma - \Sigma\tilde{\gamma}^{NM}\right) \\ \Leftrightarrow \quad & \tilde{\gamma}^{NM} = -\frac{1-\Sigma}{1+\Sigma}\mu_\gamma. \end{aligned}$$

As $\tilde{\gamma}^{SV} = -\frac{1-\Sigma}{\Sigma}\mu_\gamma$ we can compute $\tilde{\gamma}^{NM} - \tilde{\gamma}^{SV}$ as

$$\tilde{\gamma}^{NM} - \tilde{\gamma}^{SV} = \frac{1-\Sigma}{(1+\Sigma)\,\Sigma}\mu_\gamma. \tag{1.A.12}$$

By Lemma 1.5, the assumptions of the Proposition imply $m_B^* > m_A^*$. Therefore, since $c''(\cdot) > 0$,

$$c'\left(m_B^*\right) - c'\left(m_A^*\right) > 0.$$

The first-order conditions for the equilibrium manipulation intensities can be rewritten as follows:

$$g_\gamma\left(\tilde{\gamma}^*\right)\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2} - \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \tilde{\gamma}^{SV}\right)\right] = \kappa_A c'\left(m_A^*\right) \tag{1.A.13}$$

and

$$g_\gamma\left(\tilde{\gamma}^*\right)\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2} + \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \tilde{\gamma}^{SV}\right)\right] = \kappa_B c'\left(m_B^*\right). \tag{1.A.14}$$

We can solve (1.A.13) and (1.A.14) for $c'\left(m_B^*\right)$ and $c'\left(m_A^*\right)$, respectively, such that

$$\begin{aligned} & \frac{g_\gamma\left(\tilde{\gamma}^*\right)}{\kappa_B}\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2} + \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \tilde{\gamma}^{SV}\right)\right] \\ > & \frac{g_\gamma\left(\tilde{\gamma}^*\right)}{\kappa_A}\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2} - \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \tilde{\gamma}^{SV}\right)\right] \\ \Leftrightarrow \quad & \frac{\kappa_A - \kappa_B}{\Sigma\left(\kappa_A + \kappa_B\right)}\left[2b + \bar{q}_{B0} + \bar{q}_{A0}\right] > \tilde{\gamma}^{SV} - \tilde{\gamma}^*. \end{aligned}$$

Hence, a sufficient condition for

$$\tilde{\gamma}^{NM} - \tilde{\gamma}^{SV} > \tilde{\gamma}^{SV} - \tilde{\gamma}^*$$

is given by

$$\tilde{\gamma}^{NM} - \tilde{\gamma}^{SV} \geq \frac{\kappa_A - \kappa_B}{\Sigma\left(\kappa_A + \kappa_B\right)}\left[2b + \bar{q}_{B0} + \bar{q}_{A0}\right].$$

By using (1.A.12), this yields

$$\frac{1-\Sigma}{(1+\Sigma)\,\Sigma}\mu_\gamma \;\geq\; \frac{\kappa_A-\kappa_B}{\Sigma\left(\kappa_A+\kappa_B\right)}\left[2b+\bar{q}_{B0}+\bar{q}_{A0}\right].$$

Plugging in for μ_γ and collecting all $\bar{q}_{B0}$ and $\bar{q}_{A0}$ on the left-hand side, gives

$$\bar{q}_{B0}\left(\frac{1-\Sigma}{1+\Sigma}-\frac{\kappa_A-\kappa_B}{\kappa_A+\kappa_B}\right)-\bar{q}_{A0}\left(\frac{1-\Sigma}{1+\Sigma}+\frac{\kappa_A-\kappa_B}{\kappa_A+\kappa_B}\right) \;\geq\; 2b\frac{\kappa_A-\kappa_B}{\kappa_A+\kappa_B},$$

which is condition (1.19) of Proposition 1.4.

Note that the upper bound of $\tilde{\gamma}^{SV}-\tilde{\gamma}^{*}$ is independent of α, but $\tilde{\gamma}^{NM}-\tilde{\gamma}^{SV}$ depends on α, because $\tilde{\gamma}^{NM}$ depends on α. From the proof of Lemma 1.1, we know that $\frac{\partial\tilde{\gamma}^{NM}}{\partial\alpha}>0$, which concludes the proof that $\mathscr{E}\left(\tilde{\gamma}^{*}\right)>\mathscr{E}\left(\tilde{\gamma}^{NM}\right)$ under the conditions of Proposition 1.4.

To sum up, we have $\mathscr{E}(\tilde{\gamma}^{*})>\mathscr{E}(\tilde{\gamma}^{NM})$. However, $m_B^{*}\neq m_A^{*}$ implies $\mathscr{P}\left(m_A^{*},m_B^{*}\right)>\mathscr{P}^{NM}$. Hence, there exists a critical value $\beta=\beta^{**}$ such that $\mathscr{W}^{*}=\mathscr{W}^{NM}$, and $\mathscr{W}^{*}\gtrless\mathscr{W}^{NM}$ if $\beta\lessgtr\beta^{**}$. □

Proof of Proposition 1.5. In the following, we show that, under the conditions of the Proposition, the cutoff ranking

$$\tilde{\gamma}^{SV}\;<\;\tilde{\gamma}^{NM}\;<\;\tilde{\gamma}^{*}$$

holds. By Lemmas 1.1 and 1.4, it suffices to show that condition (1.20) implies $m_A^{*}>m_B^{*}$. We begin the proof by finding a necessary condition for $m_A^{*}\leq m_B^{*}$. According to (1.A.13) and (1.A.14), the condition $m_A^{*}\leq m_B^{*}$ will hold if and only if

$$\begin{aligned}
&\frac{g_\gamma\left(\tilde{\gamma}^{*}\right)}{\kappa_A}\left[b+\frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}-\frac{\Sigma}{2}\left(\tilde{\gamma}^{*}-\tilde{\gamma}^{SV}\right)\right]\\
\leq\;&\frac{g_\gamma\left(\tilde{\gamma}^{*}\right)}{\kappa_B}\left[b+\frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}+\frac{\Sigma}{2}\left(\tilde{\gamma}^{*}-\tilde{\gamma}^{SV}\right)\right]\\
\Leftrightarrow\quad&2b\frac{\kappa_B-\kappa_A}{\Sigma\left(\kappa_A+\kappa_B\right)}+\frac{\kappa_B-\kappa_A}{\Sigma\left(\kappa_A+\kappa_B\right)}\left(\bar{q}_{A0}+\bar{q}_{B0}\right)\;\leq\;\tilde{\gamma}^{*}-\tilde{\gamma}^{SV}.
\end{aligned}$$

By Lemma 1.4, we can conclude that

$$2b\frac{\kappa_B-\kappa_A}{\Sigma\left(\kappa_A+\kappa_B\right)}+\frac{\kappa_B-\kappa_A}{\Sigma\left(\kappa_A+\kappa_B\right)}\left(\bar{q}_{A0}+\bar{q}_{B0}\right)\;\leq\;\tilde{\gamma}^{NM}-\tilde{\gamma}^{SV}$$

is a necessary condition. It follows that if

$$2b\frac{\kappa_B-\kappa_A}{\Sigma\left(\kappa_A+\kappa_B\right)}+\frac{\kappa_B-\kappa_A}{\Sigma\left(\kappa_A+\kappa_B\right)}\left(\bar{q}_{A0}+\bar{q}_{B0}\right) > \tilde{\gamma}^{NM}-\tilde{\gamma}^{SV}, \tag{1.A.15}$$

then $m_A^* > m_B^*$. Recall that we proved $\frac{\partial\tilde{\gamma}^{NM}}{\partial\alpha} > 0$ in the proof of Lemma 1.1. Additionally, note that the left-hand side of (1.A.15) and $\tilde{\gamma}^{SV}$ are independent of α. Therefore, it suffices to show that (1.A.15) is fulfilled for $\alpha = 1$. If $\alpha = 1$, then $\tilde{\gamma}^{NM} = 0$ by (1.15) and, hence, $\tilde{\gamma}^{NM}-\tilde{\gamma}^{SV} = (1-\Sigma)\mu_\gamma/\Sigma$. Plugging this into (1.A.15), yields

$$2b\frac{\kappa_B-\kappa_A}{\Sigma\left(\kappa_A+\kappa_B\right)}+\frac{\kappa_B-\kappa_A}{\Sigma\left(\kappa_A+\kappa_B\right)}\left(\bar{q}_{A0}+\bar{q}_{B0}\right) > \frac{1-\Sigma}{\Sigma}\left(\bar{q}_{B0}-\bar{q}_{A0}\right),$$

which can be rewritten to (1.20).

To sum up, the cutoff ranking $\tilde{\gamma}^{SV} < \tilde{\gamma}^{NM} < \tilde{\gamma}^*$ implies $\mathscr{E}\left(\tilde{\gamma}^*\right) < \mathscr{E}\left(\tilde{\gamma}^{NM}\right)$ according to Lemma 1.2. As $\mathscr{P}(m_A^*, m_B^*) > \mathscr{P}^{NM}$ holds because of $m_B^* \neq m_A^*$, we obtain $\mathscr{W}^* < \mathscr{W}^{NM}$ for all β. □

Proof of Proposition 1.6. The proof of the Proposition is based on the results of the following lemma:

Lemma 1.7. *The following comparative statics hold:*
(i) $\partial\tilde{\gamma}^*/\partial\kappa_A < 0$ *and* $\partial\tilde{\gamma}^*/\partial\kappa_B > 0$.
(ii) If $m_A^* > m_B^*$, *then* $\partial\mathscr{P}\left(m_A^*, m_B^*\right)/\partial\kappa_A < 0$ *and* $\partial\mathscr{P}\left(m_A^*, m_B^*\right)/\partial\kappa_B > 0$.
(iii) If $m_A^* < m_B^*$, *then* $\partial\mathscr{P}\left(m_A^*, m_B^*\right)/\partial\kappa_A > 0$ *and* $\partial\mathscr{P}\left(m_A^*, m_B^*\right)/\partial\kappa_B < 0$.

Proof of Lemma 1.7. (i) We first have to compute the determinant of the Jacobian that corresponds to the set of implicit functions

$$\begin{aligned} F_A &:= g_\gamma(\tilde{\gamma})\left(b+\bar{q}_{A0}-\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)-c_A'(m_A), \\ F_B &:= g_\gamma(\tilde{\gamma})\left(b+\bar{q}_{B0}+\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)-c_B'(m_B), \\ F_{\tilde{\gamma}} &:= \alpha\cdot G_\theta(m_B-m_A+\tilde{\gamma})+(1-\alpha)\cdot G_\theta((1-\Sigma)\mu_\gamma+\Sigma\tilde{\gamma})-\frac{1}{2}, \end{aligned} \tag{1.A.16}$$

which follow from (1.8)–(1.10). The Jacobian

$$\boldsymbol{J} := \begin{pmatrix} \frac{\partial F_A}{\partial m_A} & \frac{\partial F_A}{\partial m_B} & \frac{\partial F_A}{\partial\tilde{\gamma}} \\ \frac{\partial F_B}{\partial m_A} & \frac{\partial F_B}{\partial m_B} & \frac{\partial F_B}{\partial\tilde{\gamma}} \\ \frac{\partial F_{\tilde{\gamma}}}{\partial m_A} & \frac{\partial F_{\tilde{\gamma}}}{\partial m_B} & \frac{\partial F_{\tilde{\gamma}}}{\partial\tilde{\gamma}} \end{pmatrix} \tag{1.A.17}$$

has the determinant

$$
\begin{aligned}
|\boldsymbol{J}| &= c_A''(m_A)c_B''(m_B)\Sigma(1-\alpha)g_\theta((1-\Sigma)\mu_\gamma+\Sigma\tilde{\gamma}) \\
&\quad + \alpha g_\theta(m_B-m_A+\tilde{\gamma})\times \\
&\quad \left(-c_A''(m_A)\left\{-g_\gamma'(\tilde{\gamma})\left(b+\bar{q}_{B0}+\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)-\frac{\Sigma}{2}g_\gamma(\tilde{\gamma})-c_B''(m_B)\right\}\right. \\
&\quad \left.+c_B''(m_B)\left[\frac{\Sigma}{2}g_\gamma(\tilde{\gamma})-g_\gamma'(\tilde{\gamma})\left(b+\bar{q}_{A0}-\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)\right]\right) \qquad (1.A.18) \\
&= c_A''(m_A)c_B''(m_B)\Sigma(1-\alpha)g_\theta((1-\Sigma)\mu_\gamma+\Sigma\tilde{\gamma}) \\
&\quad + \alpha g_\theta(m_B-m_A+\tilde{\gamma})\times \\
&\quad \left(-c_B''(m_B)\left\{g_\gamma'(\tilde{\gamma})\left(b+\bar{q}_{A0}-\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)-\frac{\Sigma}{2}g_\gamma(\tilde{\gamma})-c_A''(m_A)\right\}\right. \\
&\quad \left.+c_A''(m_A)\left[g_\gamma'(\tilde{\gamma})\left(b+\bar{q}_{B0}+\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)+\frac{\Sigma}{2}g_\gamma(\tilde{\gamma})\right]\right). \qquad (1.A.19)
\end{aligned}
$$

As, by assumption, the candidates have strictly concave objective functions, the second-order conditions

$$
-g_\gamma'(\tilde{\gamma})\left(b+\bar{q}_{B0}+\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)-\frac{\Sigma}{2}g_\gamma(\tilde{\gamma})-c_B''(m_B) \;<\; 0 \qquad (1.A.20)
$$

and

$$
g_\gamma'(\tilde{\gamma})\left(b+\bar{q}_{A0}-\frac{\Sigma}{2}\left(\tilde{\gamma}-\mu_\gamma\right)\right)-\frac{\Sigma}{2}g_\gamma(\tilde{\gamma})-c_A''(m_A) \;<\; 0 \qquad (1.A.21)
$$

must hold. The expression in (1.A.18) is clearly positive if $g_\gamma'(\tilde{\gamma})=0$. If $g_\gamma'(\tilde{\gamma})<0$ it directly follows from (1.A.18) and (1.A.20) that $|\boldsymbol{J}|>0$. If, however, $g_\gamma'(\tilde{\gamma})>0$ it directly follows from (1.A.19) and (1.A.21) that $|\boldsymbol{J}|>0$ such that we conclude that the determinant of the Jacobian is strictly positive.

Applying the implicit-function theorem, by Cramer's rule we obtain

$$
\begin{aligned}
\frac{\partial\tilde{\gamma}^*}{\partial\kappa_A} &= \frac{1}{|\boldsymbol{J}|}\cdot\left|\begin{array}{ccc} \frac{\partial F_A}{\partial m_A} & \frac{\partial F_A}{\partial m_B} & -\frac{\partial F_A}{\partial \kappa_A} \\ \frac{\partial F_B}{\partial m_A} & \frac{\partial F_B}{\partial m_B} & -\frac{\partial F_B}{\partial \kappa_A} \\ \frac{\partial F_{\tilde{\gamma}}}{\partial m_A} & \frac{\partial F_{\tilde{\gamma}}}{\partial m_B} & -\frac{\partial F_{\tilde{\gamma}}}{\partial \kappa_A} \end{array}\right|_{(m_A,m_B,\tilde{\gamma})=(m_A^*,m_B^*,\tilde{\gamma}^*)} \\
&= -\frac{c_B''(m_B^*)c'(m_A^*)}{|\boldsymbol{J}|}\cdot\alpha g_\theta(m_B^*-m_A^*+\tilde{\gamma}^*) \;<\; 0.
\end{aligned}
$$

In analogy, we can compute how the equilibrium cutoff reacts to an increase of κ_B:

$$
\frac{\partial\tilde{\gamma}^*}{\partial\kappa_B} = \frac{c_A''(m_A^*)c'(m_B^*)}{|\boldsymbol{J}|}\cdot\alpha g_\theta(m_B^*-m_A^*+\tilde{\gamma}^*) \;>\; 0. \qquad (1.A.22)
$$

(ii)–(iii) Implicit differentiation of the equilibrium manipulation intensity m_A^* with respect to the cost parameter κ_B yields

$$\frac{\partial m_A^*}{\partial \kappa_B} = \frac{c'(m_B^*)\alpha g_\theta(m_B^* - m_A^* + \tilde{\gamma}^*)}{|\boldsymbol{J}|} \times \left[g_\gamma'(\tilde{\gamma}^*)\left(b + \bar{q}_{A0} - \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right)\right) - \frac{\Sigma}{2} g_\gamma(\tilde{\gamma}^*)\right].$$

In addition,

$$\begin{aligned}\frac{\partial m_B^*}{\partial \kappa_B} = \frac{c'(m_B^*)}{|\boldsymbol{J}|} \cdot \{ & \alpha g_\theta(m_B^* - m_A^* + \tilde{\gamma}^*) \times \\ & \left[g_\gamma'(\tilde{\gamma}^*)\left(b + \bar{q}_{A0} - \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right)\right) - \frac{\Sigma}{2} g_\gamma(\tilde{\gamma}^*) - c_A''(m_A^*)\right] \\ & - c_A''(m_A^*) \cdot \Sigma(1-\alpha) g_\theta((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^*)\},\end{aligned}$$

which is negative as the expression in square brackets is negative according to (1.A.21). Intuitively, the lower a candidate's cost parameter, the higher will be his equilibrium manipulation intensity. Thus, as the net effect of κ_B on $\partial m_A^*/\partial \kappa_B - \partial m_B^*/\kappa_B$ we obtain

$$\begin{aligned}\frac{\partial m_A^*}{\partial \kappa_B} - \frac{\partial m_B^*}{\partial \kappa_B} = \frac{c'(m_B^*)c_A''(m_A^*)}{|\boldsymbol{J}|} [& \alpha g_\theta(m_B^* - m_A^* + \tilde{\gamma}^*) \\ & + \Sigma(1-\alpha) g_\theta((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^*)],\end{aligned}$$

which is strictly positive. In analogy, one can compute

$$\begin{aligned}\frac{\partial m_A^*}{\partial \kappa_A} = \frac{c'(m_A^*)}{|\boldsymbol{J}|} \cdot \{ & \alpha g_\theta(m_B^* - m_A^* + \tilde{\gamma}^*) \times \\ & \left[-g_\gamma'(\tilde{\gamma}^*)\left(b + \bar{q}_{B0} + \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right)\right) - \frac{\Sigma}{2} g_\gamma(\tilde{\gamma}^*) - c_B''(m_B^*)\right] \\ & - c_B''(m_B^*) \cdot \Sigma(1-\alpha) g_\theta((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^*)\}\end{aligned}$$

and

$$\frac{\partial m_B^*}{\partial \kappa_A} = \frac{-c'(m_A^*)\alpha g_\theta(m_B^* - m_A^* + \tilde{\gamma}^*)}{|\boldsymbol{J}|} \times \left[g_\gamma'(\tilde{\gamma}^*)\left(b + \bar{q}_{B0} + \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right)\right) + \frac{\Sigma}{2} g_\gamma(\tilde{\gamma}^*)\right],$$

leading to

$$\frac{\partial m_A^*}{\partial \kappa_A} - \frac{\partial m_B^*}{\partial \kappa_A} = -\frac{c'(m_A^*)c_B''(m_B^*)}{|\boldsymbol{J}|} [\alpha g_\theta(m_B^* - m_A^* + \tilde{\gamma}^*)$$

$$+ \Sigma(1-\alpha) g_\theta((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^*)] < 0.$$

Hence, we get that $m_A^* - m_B^*$ is increasing in κ_B and decreasing in κ_A. As $\mathscr{P}\left(m_A^*, m_B^*\right)$ is increasing in $(m_A^* - m_B^*)^2$ this insight completes the proof. □

The proof of the Proposition consists of two steps, leading to the observations that are summarized in Figure 1.A.1. As a *first step*, we will show that for every κ_A there exists an unique $\underline{\kappa}_B\left(\kappa_A\right)$ with $\underline{\kappa}_B\left(\kappa_A\right) < \kappa_A$ such that $\tilde{\gamma}^* = \tilde{\gamma}^{SV}$ if $\kappa_B = \underline{\kappa}_B\left(\kappa_A\right)$. At the same time, we show a similar claim for κ_B: For every κ_B there exists an unique $\bar{\kappa}_A\left(\kappa_B\right)$ with $\bar{\kappa}_A\left(\kappa_B\right) > \kappa_B$ such that $\tilde{\gamma}^* = \tilde{\gamma}^{SV}$ if $\kappa_A = \bar{\kappa}_A\left(\kappa_B\right)$. Consider the equilibrium cutoff candidate $\tilde{\gamma}^{SV}$. It will be part of an equilibrium if and only if the equilibrium conditions as defined in Proposition 1.1 are fulfilled, i.e., if and only if there exist m_i^{SV} and m_j^{SV} with $i,j \in \{A,B\}$ and $i \neq j$ such that

$$g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2}\right] = \kappa_i c'\left(m_i^{SV}\right) \tag{1.A.23}$$

$$g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2}\right] = \kappa_j c'\left(m_j^{SV}\right)$$

$$\frac{1}{2} = \alpha \cdot G_\theta(m_B^{SV} - m_A^{SV} + \tilde{\gamma}^{SV}) + (1-\alpha) \cdot G_\theta(0). \tag{1.A.24}$$

For every κ_i, the manipulation intensity m_i^{SV} is uniquely defined by (1.A.23). Furthermore, (1.A.24) implies

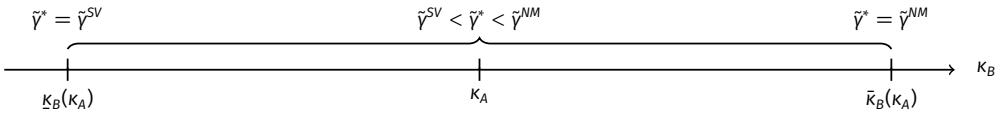
$$m_B^{SV} - m_A^{SV} + \tilde{\gamma}^{SV} = 0 \quad \Leftrightarrow \quad m_B^{SV} = \frac{1-\Sigma}{\Sigma}\mu_\gamma + m_A^{SV}.$$

Consider $i = A$ first. Consequently, $\tilde{\gamma}^{SV}$ will be an equilibrium cutoff if and only if

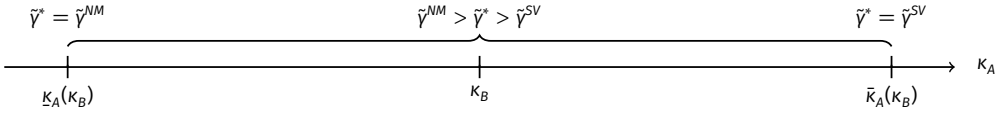
$$g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2}\right] = \kappa_B c'\left(\frac{1-\Sigma}{\Sigma}\mu_\gamma + m_A^{SV}\right).$$

The argument of the marginal cost function is independent of κ_B. Hence, for

$$\kappa_B = \underline{\kappa}_B\left(\kappa_A\right) := \frac{g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}\right]}{c'\left(\frac{1-\Sigma}{\Sigma}\mu_\gamma + m_A^{SV}\right)} > 0$$



(a) The interplay between the individual cost parameter of candidate A (κ_A) and the equilibrium cutoff ($\tilde{\gamma}^*$).



(b) The interplay between the individual cost parameter of candidate B (κ_B) and the equilibrium cutoff ($\tilde{\gamma}^*$).

Figure 1.A.1. The interplay between the candidates' individual cost parameters (κ_A, κ_B) and the equilibrium cutoff ($\tilde{\gamma}^*$).

the equilibrium cutoff equals the sophisticated-voting cutoff. Note that $\underline{\kappa}_B\left(\kappa_A\right)$ is unique due to Lemma 1.7(i). We have claimed that $\kappa_A > \underline{\kappa}_B\left(\kappa_A\right)$. This holds true if and only if

$$\frac{g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}\right]}{c'\left(m_A^{SV}\right)} > \frac{g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}\right]}{c'\left(\frac{1-\Sigma}{\Sigma}\mu_\gamma + m_A^{SV}\right)},$$

which is satisfied as $(1-\Sigma)\mu_\gamma/\Sigma > 0$.

Consider $i = B$. $\tilde{\gamma}^{SV}$ will be an equilibrium cutoff if and only if

$$g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0} + \bar{q}_{B0}}{2}\right] = \kappa_A c'\left(m_B^{SV} - \frac{1-\Sigma}{\Sigma}\mu_\gamma\right).$$

The argument of the marginal cost function is independent of κ_A. Hence, for

$$\kappa_A = \bar{\kappa}_A\left(\kappa_B\right) := \frac{g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}\right]}{c'\left(m_B^{SV} - \frac{1-\Sigma}{\Sigma}\mu_\gamma\right)} > 0$$

the equilibrium cutoff equals the sophisticated-voting cutoff. Note that $\bar{\kappa}_A$ is unique due to Lemma 1.7(i). We have claimed that $\kappa_B < \bar{\kappa}_A\left(\kappa_B\right)$. This holds true if and only if

$$\frac{g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}\right]}{c'\left(m_B^{SV}\right)} < \frac{g_\gamma\left(\tilde{\gamma}^{SV}\right)\left[b + \frac{\bar{q}_{A0}+\bar{q}_{B0}}{2}\right]}{c'\left(m_B^{SV} - \frac{1-\Sigma}{\Sigma}\mu_\gamma\right)},$$

which is satisfied as $(1-\Sigma)\mu_\gamma/\Sigma > 0$.

As a *second step*, we will derive the two thresholds $\bar{\kappa}_B\left(\kappa_A\right)$ and $\underline{\kappa}_A\left(\kappa_B\right)$, and prove the two results of Proposition 1.6. We start with result (ii):

(ii) Fix κ_A. If $\kappa_B > \underline{\kappa}_B(\kappa_A)$, Lemma 1.7(i) and the result from our first step together imply $\tilde{\gamma}^* > \tilde{\gamma}^{SV}$. Therefore, Lemma 1.7(i) and Lemma 1.2 lead to $\partial \mathscr{E}(\tilde{\gamma}^*)/\partial \kappa_B < 0$. By Lemma 1.7(i), Proposition 1.3, the proof of Proposition 1.5, and the continuity of the equilibrium conditions, there exists a threshold which fulfills $\bar{\kappa}_B(\kappa_A) > \kappa_A$ such that $\tilde{\gamma}^* < \tilde{\gamma}^{NM}$ if and only if $\kappa_B < \bar{\kappa}_B(\kappa_A)$. By Lemma 1.4 we can infer that $m_A^* < m_B^*$ as long as $\kappa_B < \bar{\kappa}_B(\kappa_A)$ and, therefore, $\partial \mathscr{P}(m_A^*, m_B^*)/\partial \kappa_B < 0$ by Lemma 1.7(iii). Thus, $\partial \mathscr{W}^*/\partial \kappa_B > 0$ if β is sufficiently large.

Fix κ_B, now. If $\kappa_A < \bar{\kappa}_A(\kappa_B)$, the result from our first step and Lemma 1.7(i) together imply $\tilde{\gamma}^* > \tilde{\gamma}^{SV}$. Therefore, Lemma 1.7(i) and Lemma 1.2 lead to $\partial \mathscr{E}(\tilde{\gamma}^*)/\partial \kappa_A > 0$. By Lemma 1.7(i), Proposition 1.3, Proposition 1.5, and the continuity of the equilibrium conditions, there exists a threshold $\underline{\kappa}_A(\kappa_B) < \kappa_B$ such that $\tilde{\gamma}^* < \tilde{\gamma}^{NM}$ if and only if $\kappa_A > \underline{\kappa}_A(\kappa_B)$. By Lemma 1.4 we can infer that $m_A^* < m_B^*$ as long as $\kappa_A > \underline{\kappa}_A(\kappa_B)$ and, therefore, $\partial \mathscr{P}(m_A^*, m_B^*)/\partial \kappa_A > 0$ by Lemma 1.7(iii). Hence, $\partial \mathscr{W}^*/\partial \kappa_A > 0$ if β is sufficiently small.

We divide the proof of result (i) in two parts: (ia) Fix κ_A. From the first step we know that $\tilde{\gamma}^* = \tilde{\gamma}^{SV}$ for $\kappa_B = \underline{\kappa}_B(\kappa_A)$ with $\underline{\kappa}_B(\kappa_A) < \kappa_A$. Lemma 1.7(i) implies that for all $\kappa_B < \underline{\kappa}_B(\kappa_A)$ we have $\tilde{\gamma}^* < \tilde{\gamma}^{SV}$ and, therefore, $\partial \mathscr{E}(\tilde{\gamma}^*)/\partial \kappa_B > 0$ for all $\kappa_B < \underline{\kappa}_B(\kappa_A)$ by Lemma 1.2. Furthermore, by Lemma 1.4, $m_A^* < m_B^*$ because $\tilde{\gamma}^* < \tilde{\gamma}^{SV}$ implies $\tilde{\gamma}^* < \tilde{\gamma}^{NM}$ by Lemma 1.1. Then, Lemma 1.7(iii) implies $\partial \mathscr{P}(m_A^*, m_B^*)/\partial \kappa_B < 0$. As $\kappa_B < \underline{\kappa}_B(\kappa_A) < \kappa_A$, κ_B increases if it moves towards κ_A and, therefore, $\mathscr{E}(\tilde{\gamma}^*)$ increases and $\mathscr{P}(m_A^*, m_B^*)$ decreases if κ_B moves towards κ_A.

Note that $\kappa_B = \bar{\kappa}_B(\kappa_A)$ implies $\tilde{\gamma}^* = \tilde{\gamma}^{NM}$. If $\kappa_B > \bar{\kappa}_B(\kappa_A)$, the cutoff ranking $\tilde{\gamma}^{SV} < \tilde{\gamma}^{NM} < \tilde{\gamma}^*$ will hold such that Lemma 1.7(i) and Lemma 1.2 lead to $\partial \mathscr{E}(\tilde{\gamma}^*)/\partial \kappa_B < 0$. Additionally, $m_A^* > m_B^*$ by Lemma 1.4. By Lemma 1.7(ii), this implies $\partial \mathscr{P}(m_A^*, m_B^*)/\partial \kappa_B > 0$. As $\kappa_B > \bar{\kappa}_B(\kappa_A) > \kappa_A$, κ_B decreases if it moves towards κ_A and, therefore, $\mathscr{E}(\tilde{\gamma}^*)$ increases and $\mathscr{P}(m_A^*, m_B^*)$ decreases if κ_B moves towards κ_A.

(ib) Fix κ_B now. From the first step we know that $\tilde{\gamma}^* = \tilde{\gamma}^{SV}$ for $\kappa_A = \bar{\kappa}_A(\kappa_B)$ with $\bar{\kappa}_A(\kappa_B) > \kappa_B$. Lemma 1.7(i) implies that for all $\kappa_A > \bar{\kappa}_A(\kappa_B)$ we have $\tilde{\gamma}^* < \tilde{\gamma}^{SV}$ and, therefore, $\partial \mathscr{E}(\tilde{\gamma}^*)/\partial \kappa_A < 0$ for all $\kappa_A > \bar{\kappa}_A(\kappa_B)$ by Lemma 1.2. Furthermore, by Lemma 1.4, $m_A^* < m_B^*$ because $\tilde{\gamma}^* < \tilde{\gamma}^{SV}$ implies $\tilde{\gamma}^* < \tilde{\gamma}^{NM}$ by Lemma 1.1. Then,

Lemma 1.7(iii) implies $\partial \mathscr{P}\left(m_A^*, m_B^*\right)/\partial \kappa_A > 0$. As $\kappa_A > \bar{\kappa}_A\left(\kappa_B\right) > \kappa_B$, κ_A decreases if it moves towards κ_B and, therefore, $\mathscr{E}(\tilde{\gamma}^*)$ increases and $\mathscr{P}\left(m_A^*, m_B^*\right)$ decreases if κ_A moves towards κ_B.

Note that $\kappa_A = \underline{\kappa}_A\left(\kappa_B\right)$ implies $\tilde{\gamma}^* = \tilde{\gamma}^{NM}$. If $\kappa_A < \underline{\kappa}_A\left(\kappa_B\right)$, the cutoff ranking $\tilde{\gamma}^{SV} < \tilde{\gamma}^{NM} < \tilde{\gamma}^*$ will hold such that Lemma 1.7(i) and Lemma 1.2 lead to $\partial \mathscr{E}(\tilde{\gamma}^*)/\partial \kappa_A > 0$. Additionally, $m_A^* > m_B^*$ by Lemma 1.4. By Lemma 1.7(ii), this implies $\partial \mathscr{P}\left(m_A^*, m_B^*\right)/\partial \kappa_A < 0$. As $\kappa_A < \underline{\kappa}_A\left(\kappa_B\right) < \kappa_B$, κ_A increases if it moves towards κ_B and, therefore, $\mathscr{E}(\tilde{\gamma}^*)$ increases and $\mathscr{P}\left(m_A^*, m_B^*\right)$ decreases if κ_A moves towards κ_B. To conclude, $\mathscr{W}^*$ will increase for all β if cost heterogeneity becomes smaller. □

Proposition 1.7. *There exists a unique equilibrium if* $c'(m_1)/c'(m_2) \geq c'(m_3)/c'(m_4)$ *whenever* $m_1, m_2, m_3, m_4 \geq 0$ *and* $m_1 - m_2 > m_3 - m_4$.

Proof. First assume that there exist two equilibria with the same cutoff $\tilde{\gamma}^* = \tilde{\gamma}^{**}$, but different manipulation intensities. As we have shown in Proposition 1.1 that all equilibria are interior, the equilibrium manipulation intensities are described by the first-order conditions given by (1.8) and (1.9). As $c_j'(\cdot)$ is strictly increasing for all $j \in \{A, B\}$ we get $m_A^* = m_A^{**}$ and $m_B^* = m_B^{**}$ and, hence, both equilibria are identical.

Now, suppose that there exist two equilibria with heterogeneous cutoffs, and, without loss of generality, assume $\tilde{\gamma}^* < \tilde{\gamma}^{**}$. Note that if both cutoffs are part of an equilibrium, (1.7) will hold for both such that

$$\begin{aligned}\frac{1}{2} &= \alpha \cdot G_\theta(m_B^* - m_A^* + \tilde{\gamma}^*) + (1-\alpha) \cdot G_\theta((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^*) \\ &= \alpha \cdot G_\theta(m_B^{**} - m_A^{**} + \tilde{\gamma}^{**}) + (1-\alpha) \cdot G_\theta((1-\Sigma)\mu_\gamma + \Sigma\tilde{\gamma}^{**}).\end{aligned}$$

Since $\tilde{\gamma}^* < \tilde{\gamma}^{**}$ implies

$$\begin{aligned}&(1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^* \;<\; (1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{**} \\ \Leftrightarrow\; &G_\theta\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^*\right) \;<\; G_\theta\left((1-\Sigma)\,\mu_\gamma + \Sigma\tilde{\gamma}^{**}\right),\end{aligned}$$

they can only be part of two equilibria at the same time if

$$\begin{aligned}&G_\theta(m_B^* - m_A^* + \tilde{\gamma}^*) \;>\; G_\theta(m_B^{**} - m_A^{**} + \tilde{\gamma}^{**}) \\ \Leftrightarrow\; &m_B^* - m_A^* + \tilde{\gamma}^* \;>\; m_B^{**} - m_A^{**} + \tilde{\gamma}^{**}\end{aligned}$$

$$\Leftrightarrow \left(m_B^* - m_A^*\right) - \left(m_B^{**} - m_A^{**}\right) > \tilde{\gamma}^{**} - \tilde{\gamma}^* > 0$$
$$\Rightarrow m_B^* - m_A^* > m_B^{**} - m_A^{**}.$$

By Proposition 1.1 all equilibria are interior. Therefore, (1.8) and (1.9) have to hold in both equilibria such that

$$g_\gamma(\tilde{\gamma}^*)\left(b + \bar{q}_{A0} - \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right)\right) = c_A'(m_A^*)$$

as well as

$$g_\gamma(\tilde{\gamma}^*)\left(b + \bar{q}_{B0} + \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \mu_\gamma\right)\right) = c_B'(m_B^*),$$

and

$$g_\gamma(\tilde{\gamma}^{**})\left(b + \bar{q}_{A0} - \frac{\Sigma}{2}\left(\tilde{\gamma}^{**} - \mu_\gamma\right)\right) = c_A'(m_A^{**})$$

as well as

$$g_\gamma(\tilde{\gamma}^{**})\left(b + \bar{q}_{B0} + \frac{\Sigma}{2}\left(\tilde{\gamma}^{**} - \mu_\gamma\right)\right) = c_B'(m_B^{**}).$$

Dividing the first-order conditions for candidate A by $g_\gamma(\tilde{\gamma}^*)$ and $g_\gamma(\tilde{\gamma}^{**})$, respectively, and subtracting one from the other yields

$$\frac{\kappa_A c'\left(m_A^*\right)}{g_\gamma\left(\tilde{\gamma}^*\right)} - \frac{\kappa_A c'\left(m_A^{**}\right)}{g_\gamma\left(\tilde{\gamma}^{**}\right)} = \frac{\Sigma}{2}\left(\tilde{\gamma}^{**} - \tilde{\gamma}^*\right) > 0,$$

such that

$$\frac{c'\left(m_A^*\right)}{c'\left(m_A^{**}\right)} > \frac{g_\gamma\left(\tilde{\gamma}^*\right)}{g_\gamma\left(\tilde{\gamma}^{**}\right)}.$$

Analogous steps for candidate B yield

$$\frac{\kappa_B c'\left(m_B^*\right)}{g_\gamma\left(\tilde{\gamma}^*\right)} - \frac{\kappa_B c'\left(m_B^{**}\right)}{g_\gamma\left(\tilde{\gamma}^{**}\right)} = \frac{\Sigma}{2}\left(\tilde{\gamma}^* - \tilde{\gamma}^{**}\right) < 0,$$

such that

$$\frac{c'\left(m_B^*\right)}{c'\left(m_B^{**}\right)} < \frac{g_\gamma\left(\tilde{\gamma}^*\right)}{g_\gamma\left(\tilde{\gamma}^{**}\right)}.$$

Taking both inequalities together we get

$$\frac{c'\left(m_B^*\right)}{c'\left(m_B^{**}\right)} < \frac{c'\left(m_A^*\right)}{c'\left(m_A^{**}\right)} \Leftrightarrow \frac{c'\left(m_B^*\right)}{c'\left(m_A^*\right)} < \frac{c'\left(m_B^{**}\right)}{c'\left(m_A^{**}\right)}.$$

However, by the assumption about the cost function stated in the Proposition, it holds for all m_B^*, m_A^*, m_B^{**}, m_A^{**} with $m_B^* - m_A^* > m_B^{**} - m_A^{**}$ that

$$\frac{c'\left(m_B^*\right)}{c'\left(m_A^*\right)} \geq \frac{c'\left(m_B^{**}\right)}{c'\left(m_A^{**}\right)},$$

which yields a contradiction and concludes the proof of the Proposition. □

The Proposition derives a condition on the cost function such that uniqueness is ensured. This restriction holds, for example, for every member of the class of exponential cost functions $c(m) = a^{\lambda m} - 1$ with $a, \lambda > 0$ and $a \neq 1$. Assuming exponential cost functions is common in the literature; see, e.g., Tadelis (2002), Olson and Roy (2008), Gershkov and Perry (2009), Picard and Tabuchi (2010), DellaVigna and Pope (2017), and DellaVigna (2018).

Appendix 1.B Micro-Foundation for the Welfare Loss from Political Polarization

In this section, we depart from the basic model and consider a modified game that evolves according to the following four steps. First, candidate $j \in \{A, B\}$ chooses manipulation intensity m_j and incurs costs $c_j\left(m_j\right)$. Second, public signals realize, and sophisticated voters update their prior beliefs, whereas naive voters take the two signals at face value. Third, voters cast their ballot according to sincere voting, and the winner of the election is chosen via simple majority rule. Fourth, voters engage in political disputes. As the first three steps are identical with the basic model, in the following we will focus on step four, where $2n$ voters interact in n social conflicts with n being arbitrarily large. After having solved for the equilibrium at the conflict stage, we will adapt the welfare measure to our modified game and offer a micro-foundation for the welfare loss from political polarization that is based on social conflict.

Suppose that, after the election, voters interact in their daily life (e.g., as colleagues, neighbors, or gym members). During these interactions, they become involved in political disputes wherein they can spend resources to bring forward their line of argument and to defend their opinion. Depending on the extent to which voters are willing to engage in such disputes, resource spending can range from investing time in pleasant conversations, over quarreling and heated debates, up to the usage of physical violence or even committing hate crime. To model such interactions, we draw on a modified version of the contest models that have been employed to study social conflict (see, among others, Esteban and Ray, 1999, 2011). After the election, $n \in \mathbb{N}$ groups of voters are formed. Each group consists of two randomly drawn voters from the electorate. Within their group, the two voters observe the political attitude of their opponent and quarrel over the competence of the two candidates. We assume that giving in (prevailing) in a political dispute against an opponent is more distressing (joyful), the more the political attitudes of the group members differ. More specifically, giving in in a group consisting of the voters 1 and 2 with ideologies $\theta_1, \theta_2 \in \mathbb{R}$, and sophistication types $i_1, i_2 \in \{S, N\}$ leads to a utility loss of

$-(\Delta u_{\theta_1}^{i_1} - \Delta u_{\theta_2}^{i_2})^2$, whereas prevailing induces a utility gain of $(\Delta u_{\theta_1}^{i_1} - \Delta u_{\theta_2}^{i_2})^2$.[25] To stir the dispute in his favored direction each voter spends resources $r > 0$ that come at cost $k(r) > 0$, which is assumed to be an increasing function and homogeneous of degree $K \geq 1$. The probability that voter 1 wins the dispute if he spends resources r_1 and his opponent resources r_2 is given by $r_1/(r_1 + r_2)$.

We know from the basic model that there exists an equilibrium at the election stage such that candidates' manipulation intensities are given by m_A^* and m_B^*, sophisticated voters form correct beliefs, naive voters take the received signals at face value, and all voters vote sincerely. Then the resulting political attitudes of a sophisticated and a naive voter with ideology θ, after γ has been realized, are given by (1.5) and (1.6) with $m_A = m_A^*$, $m_B = m_B^*$, $\hat{m}_A = m_A^*$ and $\hat{m}_B = m_B^*$ such that

$$\Delta u_\theta^{S*} = \theta + \bar{q}_{A1}^{S\,*} - \bar{q}_{B1}^{S\,*} = \theta - (1 - \Sigma)\mu_\gamma - \Sigma\gamma$$

and

$$\Delta u_\theta^{N*} = \theta + \bar{q}_{A1}^{N\,*} - \bar{q}_{B1}^{N\,*} = \theta - m_B^* + m_A^* - \gamma.$$

Hence, in any interior equilibrium the two exemplary voters 1 and 2 with ideologies θ_1 and θ_2, and sophistication types i_1 and i_2 simultaneously choose their resource expenditures r_1 and r_2 to maximize their objective functions

$$(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2 \cdot \frac{r_1}{r_1 + r_2} - (\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2 \cdot \frac{r_2}{r_1 + r_2} - k(r_1)$$

for voter 1, and

$$(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2 \cdot \frac{r_2}{r_1 + r_2} - (\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2 \cdot \frac{r_1}{r_1 + r_2} - k(r_2)$$

for voter 2. As both functions are strictly concave, optimal resource expenditures, r_1^* and r_2^*, are described by the respective first-order conditions, leading to

$$\frac{2(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2}{\left(r_1^* + r_2^*\right)^2} = \frac{k'(r_1^*)}{r_2^*} = \frac{k'(r_2^*)}{r_1^*}.$$

25. To keep the setting tractable, we use a quadratic function to model gains and losses at the conflict stage. Quadratic loss functions are often used in papers on political economy in order to model single-peaked preferences; see, among others, Baron (1994), Dewan and Myatt (2008), Levy and Razin (2015), Ortoleva and Snowberg (2015), and Little (2017).

The equation $r_1^* k'(r_1^*) = r_2^* k'(r_2^*)$ implies a unique symmetric equilibrium with $r_1^* = r_2^* = r^*$ being implicitly described by

$$r^* k'(r^*) \;=\; \frac{(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2}{2}. \tag{1.B.1}$$

Next, we consider the ex-ante expected utilitarian welfare in our modified game. To avoid technical issues due to the existence of infinitely many voters, we analyze welfare of the (potentially large) random subgroup of $2n$ voters that quarrel over political issues:

$$\mathscr{W} \;:=\; 2nE[u_\theta] + \beta E\left[-\sum_{l=1}^{2n} k(r_l)\right].$$

Again, welfare consists of two components. As in the basic model, every voter receives utility from the elected candidate's administration, which describes the first component. The second component, however, differs from that in the basic model. Now, voters on average incur disutility from social conflicts. While the returns from the conflict sum up to zero in each group, both players spend costly resources during the conflict. The corresponding expected costs are reflected in the second part of welfare, $E\left[-\sum_{l=1}^{2n} k(r_l)\right]$.

Above, we have shown that the conflict game between two randomly selected voters 1 and 2 has a symmetric equilibrium $r_1^* = r_2^* = r^*$ being described by (1.B.1). Thus, at the conflict stage, each voter's expected utility in equilibrium amounts to $-k(r^*)$. As the cost function $k(\cdot)$ is homogeneous of degree $K \geq 1$, the costs for spending resources r^* directly follow from (1.B.1):[26]

$$r^* k'(r^*) \;=\; \frac{(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2}{2} \;\Leftrightarrow\; k(r^*) \;=\; \frac{(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2}{2K}.$$

The part $E\left[-\sum_{l=1}^{2n} k(r_l)\right]$ of the welfare function can, thus, be computed as

$$-\frac{n}{K} \cdot E[(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2],$$

26. Homogeneity of degree K means that $k(tr_i) = t^K k(r_i)$ for all $t > 0$. Thus, we have $k(r_i) = r_i^K k(1)$ so that $k'(r_i) = K r_i^{K-1} k(1)$ and $r^* k'(r^*) = r^* K \cdot (r^*)^{K-1} k(1) = K \cdot k(r^*)$.

where the expectation operator refers to the two independent random draws $\Delta u_{\theta_1}^{i_1\,*}$ and $\Delta u_{\theta_2}^{i_2\,*}$, and the composed random variable γ. Let $h(\Delta u|\gamma)$ denote the density of political attitudes given the realization of γ, with corresponding mean $\mu_{\Delta u}$ and variance $\sigma^2_{\Delta u}$. As $\Delta u_{\theta_1}^{i_1\,**}$ and $\Delta u_{\theta_2}^{i_2\,*}$ are independently drawn from the whole electorate, $E[(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^{2^*}]$ can be computed as follows:

$$
\begin{aligned}
& E[(\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\Delta u_{\theta_1}^{i_1\,*} - \Delta u_{\theta_2}^{i_2\,*})^2 h(\Delta u_{\theta_1}^{i_1\,*}|\gamma) h(\Delta u_{\theta_2}^{i_2\,*}|\gamma) g_\gamma(\gamma) \mathrm{d}\Delta u_{\theta_1}^{i_1\,*} \mathrm{d}\Delta u_{\theta_2}^{i_2\,*} \mathrm{d}\gamma \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [(\Delta u_{\theta_1}^{i_1\,*})^2 - 2\Delta u_{\theta_1}^{i_1\,*} \Delta u_{\theta_2}^{i_2\,*} + (\Delta u_{\theta_2}^{i_2\,*})^2] \\
&\qquad \cdot h(\Delta u_{\theta_1}^{i_1\,*}|\gamma) h(\Delta u_{\theta_2}^{i_2\,*}|\gamma) g_\gamma(\gamma) \mathrm{d}\Delta u_{\theta_1}^{i_1\,*} \mathrm{d}\Delta u_{\theta_2}^{i_2\,*} \mathrm{d}\gamma \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [(\Delta u_{\theta_1}^{i_1\,*})^2 - 2\Delta u_{\theta_1}^{i_1\,*} \mu_{\Delta u} + 2\Delta u_{\theta_1}^{i_1\,*} \mu_{\Delta u} + \mu^2_{\Delta u} - \mu^2_{\Delta u} + (\Delta u_{\theta_2}^{i_2\,*})^2 \\
&\qquad - 2\Delta u_{\theta_2}^{i_2\,*} \mu_{\Delta u} + 2\Delta u_{\theta_2}^{i_2\,*} \mu_{\Delta u} + \mu^2_{\Delta u} - \mu^2_{\Delta u} - 2\Delta u_{\theta_1}^{i_1\,*} \Delta u_{\theta_2}^{i_2\,*}] \\
&\qquad \cdot h(\Delta u_{\theta_1}^{i_1\,*}|\gamma) h(\Delta u_{\theta_2}^{i_2\,*}|\gamma) g_\gamma(\gamma) \mathrm{d}\Delta u_{\theta_1}^{i_1\,*} \mathrm{d}\Delta u_{\theta_2}^{i_2\,*} \mathrm{d}\gamma \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [(\Delta u_{\theta_1}^{i_1\,*} - \mu_{\Delta u})^2 + (\Delta u_{\theta_2}^{i_2\,*} - \mu_{\Delta u})^2 + 2\Delta u_{\theta_1}^{i_1\,*} \mu_{\Delta u} + 2\Delta u_{\theta_2}^{i_2\,*} \mu_{\Delta u} \\
&\qquad - \mu^2_{\Delta u} - \mu^2_{\Delta u} - 2\Delta u_{\theta_1}^{i_1\,*} \Delta u_{\theta_2}^{i_2\,*}] \\
&\qquad \cdot h(\Delta u_{\theta_1}^{i_1\,*}|\gamma) h(\Delta u_{\theta_2}^{i_2\,*}|\gamma) g_\gamma(\gamma) \mathrm{d}\Delta u_{\theta_1}^{i_1\,*} \mathrm{d}\Delta u_{\theta_2}^{i_2\,*} \mathrm{d}\gamma \\
&= 2\int_{-\infty}^{\infty} \sigma^2_{\Delta u} g_\gamma(\gamma) \mathrm{d}\gamma.
\end{aligned}
$$

Let $h^i(\Delta u_\theta^{i\,*}|\gamma)$ denote the density of political attitudes for voters of type i given the realization of γ. Thus,

$$h\left(\Delta u|\gamma\right) = \alpha h^N\left(\Delta u|\gamma\right) + (1-\alpha)\, h^S\left(\Delta u|\gamma\right),$$

so that we can write the variance $\sigma^2_{\Delta u}$ as follows:

$$\sigma^2_{\Delta u} = \int_{-\infty}^{\infty} x^2 h(x|\gamma) dx - \mu^2_{\Delta u}$$

$$= \alpha \int_{-\infty}^{\infty} x^2 h^N(x|\gamma)dx + (1-\alpha) \int_{-\infty}^{\infty} x^2 h^S(x|\gamma)dx - \mu_{\Delta u}^2$$
$$= \alpha\left[(\bar{q}_{A1}^{N\,*} - \bar{q}_{B1}^{N\,*})^2 + \sigma_\theta^2\right] + (1-\alpha)\left[(\bar{q}_{A1}^{S\,*} - \bar{q}_{B1}^{S\,*})^2 + \sigma_\theta^2\right] - \mu_{\Delta u}^2. \tag{1.B.2}$$

The mean $\mu_{\Delta u}$ is given by

$$\mu_{\Delta u} = \int_{-\infty}^{\infty} xh\left(x|\gamma\right)dx = \int_{-\infty}^{\infty} x(\alpha h^N\left(x|\gamma\right) + (1-\alpha)h^S\left(x|\gamma\right))dx$$
$$= \alpha(\bar{q}_{A1}^{N\,*} - \bar{q}_{B1}^{N\,*}) + (1-\alpha)(\bar{q}_{A1}^{S\,*} - \bar{q}_{B1}^{S\,*}),$$

where the last equality follows from the fact that the density $h^i\left(\cdot|\gamma\right)$ has the mean $\bar{q}_{A1}^{i\,*} - \bar{q}_{B1}^{i\,*}$. Inserting for $\mu_{\Delta u}$ in (1.B.2) leads to

$$\sigma_{\Delta u}^2 = \sigma_\theta^2 + \alpha(1-\alpha)\left[(\bar{q}_{A1}^{N\,*} - \bar{q}_{B1}^{N\,*}) - (\bar{q}_{A1}^{S\,*} - \bar{q}_{B1}^{S\,*})\right]^2.$$

As $\bar{q}_{A1}^{N\,*} - \bar{q}_{B1}^{N\,*} = -m_B^* + m_A^* - \gamma$ and $\bar{q}_{A1}^{S\,*} - \bar{q}_{B1}^{S\,*} = -(1-\Sigma)\mu_\gamma - \Sigma\gamma$, we obtain

$$\sigma_{\Delta u}^2 = \sigma_\theta^2 + \alpha(1-\alpha)\left[m_A^* - m_B^* + (1-\Sigma)(\mu_\gamma - \gamma)\right]^2.$$

Computing its expectation yields

$$E[\sigma_{\Delta u}^2] = \mathscr{P}\left(m_A^*, m_B^*\right) = \sigma_\theta^2 + \alpha(1-\alpha)\left[(m_A^* - m_B^*)^2 + (1-\Sigma)^2\sigma_\gamma^2\right].$$

To sum up, in our modified game, equilibrium welfare amounts to

$$\mathscr{W}^* = 2n\left[\mathscr{E}(\tilde{\gamma}^*) - \frac{\beta}{2K} \cdot \mathscr{P}\left(m_A^*, m_B^*\right)\right]$$

with $\mathscr{E}(\tilde{\gamma}^*)$ being the same ex-ante expected utility from candidate selection as in the basic model.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of the Nazis in Prewar Germany." *Quarterly Journal of Economics* 130 (4): 1885–939. DOI: https://doi.org/10.1093/qje/qjv030. [11]

Alesina, Alberto, Reza Baqir, and William Easterly. 1999. "Public Goods and Ethnic Divisions." *Quarterly Journal of Economics* 114 (4): 1243–84. DOI: https://doi.org/10.1162/003355399556269. [5, 20]

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The Welfare Effects of Social Media." *American Economic Review* 110 (3): 629–76. DOI: https://doi.org/10.1257/aer.20190658. [11]

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–36. DOI: https://doi.org/10.1257/jep.31.2.211. [5, 6, 10, 14, 15, 22, 30]

Bar-Isaac, Heski, and Joyee Deb. 2014. "What Is a Good Reputation? Career Concerns with Heterogeneous Audiences." *International Journal of Industrial Organization* 34: 44–50. DOI: https://doi.org/10.1016/j.ijindorg.2014.02.012. [12]

Bardhan, Pranab K, and Dilip Mookherjee. 2000. "Capture and Governance at Local and National Levels." *American Economic Review* 90 (2): 135–39. DOI: https://doi.org/10.1257/aer.90.2.135. [10]

Bardhan, Pranab K., and Dilip Mookherjee. 2005. "Decentralizing Antipoverty Program Delivery in Developing Countries." *Journal of Public Economics* 89 (4): 675–704. DOI: https://doi.org/10.1016/j.jpubeco.2003.01.001. [10]

Baron, David P. 1994. "Electoral Competition with Informed and Uninformed Voters." *American Political Science Review* 88 (1): 33–47. DOI: https://doi.org/10.2307/2944880. [6, 10, 30, 58]

Benjamin, Dan, AL Bodoh-Creed, and Matthew Rabin. 2019. "Base-Rate Neglect: Foundations and Implications." *Discussion Paper*, URL: https://cutt.ly/bhQKbTw. [30]

Besley, Timothy. 2005. "Political Selection." *Journal of Economic Perspectives* 19 (3): 43–60. DOI: https://doi.org/10.1257/089533005774357761. [12]

Binder, Sarah A. 1999. "The Dynamics of Legislative Gridlock, 1947-96." *American Political Science Review* 93 (3): 519–33. DOI: https://doi.org/10.2307/2585572. [5, 20]

Dato, Simon, Andreas Grunewald, Daniel Müller, and Philipp Strack. 2017. "Expectation-Based Loss Aversion and Strategic Interaction." *Games and Economic Behavior* 104: 681–705. DOI: https://doi.org/10.1016/j.geb.2017.06.010. [35]

Debreu, Gerard. 1952. "A Social Equilibrium Existence Theorem." *Proceedings of the National Academy of Sciences* 38 (10): 886–93. DOI: https://doi.org/10.1073/pnas.38.10.886. [36]

DeGroot, Morris H. 2005. *Optimal Statistical Decisions*. Vol. 82, John Wiley & Sons. [17, 41]

DellaVigna, Stefano. 2018. *Chapter 7 - Structural Behavioral Economics*. Edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson. DOI: https://doi.org/10.1016/bs.hesbe.2018.07.005. [56]

DellaVigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics* 122 (3): 1187–234. DOI: https://doi.org/10.1162/qjec.122.3.1187. [11]

DellaVigna, Stefano, and Devin Pope. 2017. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69. DOI: https://doi.org/10.1093/restud/rdx033. [56]

Dewan, Torun, and David P. Myatt. 2008. "The Qualities of Leadership: Direction, Communication, and Obfuscation." *American Political Science Review* 102 (3): 351–68. DOI: https://doi.org/10.1017/S0003055408080234. [58]

DiMaggio, Paul, John Evans, and Bethany Bryson. 1996. "Have American's Social Attitudes Become More Polarized?" *American Journal of Sociology* 102 (3): 690–755. DOI: https://doi.org/10.1086/230995. [20]

Eliaz, Kfir, and Ran Spiegler. 2020. "A Model of Competing Narratives." *American Economic Review* 110 (12): 3786–816. DOI: https://doi.org/10.1257/aer.20191099. [35]

Enelow, James M., and Melvin J. Hinich. 1982. "Ideology, Issues, and the Spatial Theory of Elections." *American Political Science Review* 76 (3): 493–501. DOI: https://doi.org/10.2307/1963727. [6, 15]

Enke, Benjamin, and Florian Zimmermann. 2017. "Correlation Neglect in Belief Formation." *Review of Economic Studies* 86 (1): 313–32. DOI: https://doi.org/10.1093/restud/rdx081. [30]

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk–Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models." *Econometrica* 84 (3): 1093–130. DOI: https://doi.org/10.3982/ECTA12609. [16, 17]

Esteban, Joan, and Debraj Ray. 1999. "Conflict and Distribution." *Journal of Economic Theory* 87 (2): 379–415. DOI: https://doi.org/10.1006/jeth.1999.2549. [5, 20, 21, 57]

Esteban, Joan, and Debraj Ray. 2011. "Linking Conflict to Inequality and Polarization." *American Economic Review* 101 (4): 1345–74. DOI: https://doi.org/10.1257/aer.101.4.1345. [5, 20, 21, 57]

Fan, Ky. 1952. "Fixed-point and Minimax Theorems in Locally Convex Topological Linear Spaces." *Proceedings of the National Academy of Sciences* 38 (2): 121–26. DOI: https://doi.org/10.1073/pnas.38.2.121. [36]

Gehlbach, Scott, and Alberto Simpser. 2015. "Electoral Manipulation as Bureaucratic Control." *American Journal of Political Science* 59 (1): 212–24. DOI: https://doi.org/10.1111/ajps.12122. [9, 10]

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. 2011. "The Effect of Newspaper Entry and Exit on Electoral Politics." *American Economic Review* 101 (7): 2980–3018. DOI: https://doi.org/10.1257/aer.101.7.2980. [11]

Gentzkow, Matthew, Jesse M. Shapiro, and Daniel F. Stone. 2015. "Chapter 14 - Media Bias in the Marketplace: Theory." In *Handbook of Media Economics*. Edited by Simon P. Anderson, Joel Waldfogel, and David Strömberg. Vol. 1, Handbook of Media Economics. North-Holland, 623–45. DOI: https://doi.org/10.1016/B978-0-444-63685-0.00014-0. [31]

Gerber, Alan S., James G. Gimper, Donald P. Green, and Daron R. Shaw. 2011. "How Large and Long-lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105 (1): 135–50. DOI: https://doi.org/10.1017/S000305541000047X. [11]

Gerber, Alan S., Dean Karlan, and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." *American Economic Journal: Applied Economics* 1 (2): 35–52. DOI: https://doi.org/10.1257/app.1.2.35. [11]

Gershkov, Alex, and Motty Perry. 2009. "Tournaments with Midterm Reviews." *Games and Economic Behavior* 66 (1): 162–90. DOI: https://doi.org/10.1016/j.geb.2008.04.003. [56]

Glicksberg, I. L. 1952. "A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points." *Proceedings of the American Mathematical Society* 3 (1): 170–74. DOI: https://doi.org/10.2307/2032478. [36]

Gottfried, Jeffrey, and Elisa Shearer. 2016. *News Use across Social Media Platforms 2016*. URL: https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/. [6]

Grossman, Gene M., and Elhanan Helpman. 1996. "Electoral Competition and Special Interest Politics." *Review of Economic Studies* 63 (2): 265–86. DOI: https://doi.org/10.2307/2297852. [6, 10, 30, 31]

Grossman, Gene M., and Elhanan Helpman. 2001. *Special interest politics*. MIT press. [10, 31]

Grossman, Gene M., and Elhanan Helpman. 2019. "Electoral Competition with Fake News." Working Paper Series (26409): DOI: https://doi.org/10.3386/w26409. [6, 11, 15]

Grunewald, Andreas, and Matthias Kräkel. 2018. "Fake News." *IZA Discussion Paper No. 11207*, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract=3092531. [10]

Höffler, Felix, and Dirk Sliwka. 2003. "Do New Brooms Sweep Clean? When and Why Dismissing a Manager Increases The Subordinates' Performance." *European Economic Review* 47 (5): 877–90. DOI: https://doi.org/10.1016/S0014-2921(02)00272-6. [12]

Holmström, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *Review of Economic Studies* 66 (1): 169–82. DOI: https://doi.org/10.1111/1467-937X.00083. [6, 12, 13]

Inderst, Roman, and Marco Ottaviani. 2013. "Sales Talk, Cancellation Terms and the Role of Consumer Protection." *Review of Economic Studies* 80 (3): 1002–26. DOI: https://doi.org/10.1093/restud/rdt005. [30]

Jones, David R. 2001. "Party Polarization and Legislative Gridlock." *Political Research Quarterly* 54 (1): 125–41. DOI: https://doi.org/10.1177/106591290105400107. [5, 20]

Kartik, Navin, Marco Ottaviani, and Francesco Squintani. 2007. "Credulity, Lies, and Costly Talk." *Journal of Economic Theory* 134 (1): 93–116. DOI: https://doi.org/10.1016/j.jet.2006.04.003. [30]

Lazear, Edward P., and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89 (5): 841–64. DOI: https://doi.org/10.1086/261010. [36]

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. "The Science of Fake News." *Science* 359 (6380): 1094–96. DOI: https://doi.org/10.1126/science.aao2998. [5, 11, 22, 30]

Ledyard, John O. 1984. "The Pure Theory of Large Two-Candidate Elections." *Public Choice* 44 (1): 7–41. DOI: https://doi.org/10.1007/BF00124816. [6, 15]

Levy, Gilat, and Ronny Razin. 2015. "Correlation Neglect, Voting Behavior, and Information Aggregation." *American Economic Review* 105 (4): 1634–45. DOI: https://doi.org/10.1257/aer.20140134. [58]

Lindbeck, Assar, and Jörgen Weibull. 1987. "Balanced-Budget Redistribution as the Outcome of Political Competition." *Public Choice* 52 (01): 273–97. DOI: https://doi.org/10.1007/BF00116710. [6, 15]

Little, Andrew T. 2012. "Elections, Fraud, and Election Monitoring in the Shadow of Revolution." *Quarterly Journal of Political Science* 7 (3): 249–83. DOI: http://doi.org/10.1561/100.00011078. [9]

Little, Andrew T. 2017. "Propaganda and Credulity." *Games and Economic Behavior* 102: 224–32. DOI: https://doi.org/10.1016/j.geb.2016.12.006. [9, 10, 13, 30, 58]

Little, Andrew T. 2019. "The Distortion of Related Beliefs." *American Journal of Political Science* 63 (3): 675–89. DOI: https://doi.org/10.1111/ajps.12435. [17]

Maskin, Eric, and Jean Tirole. 2004. "The Politician and the Judge: Accountability in Government." *American Economic Review* 94 (4): 1034–54. DOI: https://doi.org/10.1257/0002828042002606. [14]

Medina, Alejandro. 2019. "Purchasing Certainty: Voter Accuracy and the Polarizing Effect of Campaign Contributions." *Discussion Paper*, URL: https://cutt.ly/YhQPMOD. [10, 31]

Meyer, Margaret A., and John Vickers. 1997. "Performance Comparisons and Dynamic Incentives." *Journal of Political Economy* 105 (3): 547–81. DOI: https://doi.org/10.1086/262082. [12]

Miklós-Thal, Jeanine, and Hannes Ullrich. 2014. "Belief Precision and Effort Incentives in Promotion Contests." *Economic Journal* 125 (589): 1952–63. DOI: https://doi.org/10.1111/ecoj.12162. [12]

Nalebuff, Barry J., and Joseph E. Stiglitz. 1983. "Prizes and Incentives: Towards a General Theory of Compensation and Competition." *Bell Journal of Economics* 14 (1): 21–43. DOI: https://doi.org/10.2307/3003535. [36]

Ogden, Benjamin. 2019. "The Imperfect Beliefs Voting Model." *Discussion Paper*, (08): DOI: https://doi.org/10.2139/ssrn.2431447. [30]

Oliver, J. Eric, and Thomas J. Wood. 2014. "Conspiracy Theories and the Paranoid Style(s) of Mass Opinion." *American Journal of Political Science* 58 (4): 952–66. DOI: https://doi.org/10.1111/ajps.12084. [6, 15]

Olson, Lars J., and Santanu Roy. 2008. "Controlling a Biological Invasion: A Non-Classical Dynamic Economic Model." *Economic Theory* 36: 453–69. DOI: https://doi.org/10.1007/s00199-007-0281-0. [56]

Ortoleva, Pietro, and Erik Snowberg. 2015. "Overconfidence in Political Behavior." *American Economic Review* 105 (2): 504–35. DOI: https://doi.org/10.1257/aer.20130921. [30, 58]

Ottaviani, Marco, and Francesco Squintani. 2006. "Naive Audience and Communication Bias." *International Journal of Game Theory* 35: 129–50. DOI: https://doi.org/10.1007/s00182-006-0054-1. [30]

Peisakhin, Leonid, and Arturas Rozenas. 2018. "Electoral Effects of Biased Media: Russian Television in Ukraine." *American Journal of Political Science* 62 (3): 535–50. DOI: https://doi.org/10.1111/ajps.12355. [11]

Picard, Pierre M., and Takatoshi Tabuchi. 2010. "Self-Organized Agglomerations and Transport Costs." *Economic Theory* 42: 565–89. DOI: https://doi.org/10.1007/s00199-008-0410-4. [56]

Prat, Andrea. 2002. "Campaign Advertising and Voter Welfare." *Review of Economic Studies* 69 (4): 999–1017. DOI: https://doi.org/10.1111/1467-937X.00234. [10]

Schöttner, Anja. 2008. "Fixed-Prize Tournaments Versus First-Price Auctions in Innovation Contests." *Economic Theory* 35 (1): 57–71. DOI: https://doi.org/10.1007/s00199-007-0208-9. [36]

Stein, Jeremy C. 1989. "Efficient Capital Markets, Inefficient Firms: A Model of Myopic Corporate Behavior*." *Quarterly Journal of Economics* 104 (4): 655–69. DOI: https://doi.org/10.2307/2937861. [12]

Szeidl, Adam, and Ferenc Szucs. 2017. "Media capture through favor exchange." *CEPR Discussion Paper No. DP11875*, URL: https://ssrn.com/abstract=2924736. [9]

Tadelis, Steven. 2002. "The Market for Reputations as an Incentive Mechanism." *Journal of Political Economy* 110 (4): 854–82. DOI: https://doi.org/10.1086/340781. [56]

Chapter 2

Consumer Protection or Efficiency? The Case of Partitioned Pricing*

Joint with Simon Dato and Fabian Schmitz

2.1 Introduction

Partitioned pricing refers to a firm's practice to split the price of a good or service into two or more components (Morwitz, Greenleaf, and Johnson, 1998).[1] Amplified by the increased prevalence of e-commerce, it seems to be the norm nowadays that consumers face substantial additional charges for shipping, handling, or payment methods (Mohammed, 2019). Empirical evidence documents that consumers underestimate the total price when being confronted with multiple prices (Greenleaf, Johnson, Morwitz, and Shalev, 2016; Voester, Ivens, and Leischnig, 2017), such that partitioned pricing is likely to deceive consumers into overbuying. Consequently, it has come under increased scrutiny: evaluating several possible price frames, the UK Office of Fair Trading concluded that partitioned pricing has the greatest potential to cause harm for consumers (Office of Fair Trading, 2010). Furthermore, competi-

* Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866. Financial support by the DFG, grant KR 2077/3-1, is gratefully acknowledged. Declarations of interests: none.

1. Partitioned pricing initially referred to a practice in which prices were presented simultaneously. Following Friedman (2020), a broader definition of partitioned pricing also contains pricing strategies in which consumers observe prices with a delay such as drip pricing and pricing of unavoidable add-ons as sub-cases.

tion authorities have penalized firms from several industries for having engaged in partitioned or drip pricing.[2]

For partitioned pricing to exploit consumers by affecting their purchasing decisions, it needs to hold that (i) firms are allowed to partition prices and (ii) consumers mistakenly undervalue the total price. Accordingly, authorities can and actually do engage in two types of policies that address each requirement to protect consumers. Regarding the first type, they have invoked policies that regulate extra fees or additional prices, even up to a ban of partitioned pricing. In recent legislation, consumer protection agencies have either banned (UK, USA[3], EU[4]) or limited (Australia[5]) payment surcharges for the use of credit or debit cards. In Australia, agencies have ruled that surcharges must not be excessive, i.e., above the transaction's actual costs to the merchant. Similarly, while many countries have banned fuel-surcharges in the airline industry, Japan still allows airlines to add a fuel-surcharge to their price. However, the surcharge is tied to the actual costs of fuel two months before the flight.[6] In a similar vein, several lawsuits dealing with excessively high interest rates for car loans bought at a car dealer combined with a car were settled by capping the mark-up a dealer can add to the actual interest rate (Cohen, 2012). As this first type of policy limits the firms' choice set, i.e., the set of prices to choose from, it can be labeled as a hard intervention (Heidhues and Kőszegi, 2018).

The second type of policy does not regulate or ban partitioned pricing. Instead, it aims to lower the degree to which consumers underestimate additional prices by increasing either (i) consumers' awareness or (ii) the degree of the additional prices' transparency. One possibility to increase price transparency is to restrict firms' ability to hide additional prices. The Australian Competition and Consumer Commission (ACCC), for instance, requires that whenever firms advertise prices, the total price

2. The Australian Competition and Consumer Commission imposed sanctions on the two largest airlines (https://cutt.ly/shLcMbb). Likewise, the Canadian Competition Bureau fined rental car agencies for charging hidden fees (https://cutt.ly/ohLvtl1).

3. For the UK and the USA, see https://cutt.ly/jhLilo7.

4. See, the Directive (EU) 2015/2366 of the European Parliament and of the Council https://cutt.ly/ehLiQby.

5. See, https://cutt.ly/xhLiYxm.

6. See, https://cutt.ly/AxNexGQ.

must be at least as visible as prices that do not include additional fees and charges.[7] However, the second type of policy does not necessarily intervene on the firms' side but may also educate consumers to reduce the impact of deceptive pricing strategies. For example, in 2003, the US Congress passed the *Fair and Accurate Credit Transactions Act* (FACTA), which provided better consumer financial literacy. It brought forth the *Financial Literacy and Education Commission* (FLEC), which is concerned with setting a national strategy to increase consumer awareness of credit scores and their impact on their financial decisions (Cohen, 2012). Contrary to the first type of policy, the second type of policy does not affect the set of feasible prices firms may offer. In line with the terminology used by Heidhues and Kőszegi (2018), it can be labeled as a soft intervention.

This paper analyzes the consequences of hard and soft interventions that aim at limiting the impact of partitioned pricing on consumer surplus and welfare. Since the first is arguably the measure that is mostly applied by competition authorities when designing policies, with Canada as a notable exception (Heyer, 2006), one might argue that our results on the first measure might be most relevant for practitioners. However, analyzing the effects of policy interventions on both measures is particularly relevant in light of a longstanding debate among antitrust scholars whether consumer surplus or welfare should be considered by regulators when intervening in markets. It goes back at least to Bork (1978) and was recently addressed in Wilson (2019). In particular, critics argue that using consumer surplus as a measure to evaluate policy interventions "must therefore be counted as 'distributive' to the extent that it produces outcomes that shift wealth or resources in favor of consumers even though an alternative outcome would produce greater total wealth." (Hovenkamp, 2012, p. 2472). To account for these arguments, we study the effects of policies on both measures and identify circumstances under which they lead to identical or very different policy implications.

For this purpose, we incorporate partitioned pricing and consumer naivety into the framework of Singh and Vives (1984) with differentiated products. Con-

7. See https://cutt.ly/rhLiVdp.

sequently, we allow firms to partition their prices into a headline price and an additional price. Moreover, to account for the mounting evidence that consumers underestimate partitioned prices, we assume that they are naive and underestimate the additional prices, as in, e.g., Gabaix and Laibson (2006), Chetty (2009) or Heidhues, Kőszegi, and Murooka (2016b). We model a hard intervention as imposing an upper bound on the additional price and a soft intervention as an increase in price transparency, i.e., a decrease in the degree to which consumers underestimate a given additional price. Comparative static in these parameters allows us to evaluate each policy's consequences on consumer surplus and welfare.

Our analysis reveals that firms have an incentive to charge high additional and low headline prices in markets with naive consumers and the possibility to partition prices. The reason is that naive consumers react less sensitive to additional prices than to headline prices. If the upper bound on the additional price is very high, equilibrium headline prices tend to become overly low. While very low or even negative headline prices might be feasible in some markets, they might not work well in others. For example, consumers might become suspicious and abstain from buying at all if deals look too good to be true. Likewise, in markets in which firms charge the additional price for an unavoidable add-on, arbitrage traders might force firms to charge positive prices (Heidhues, Kőszegi, and Murooka, 2016a,b). To account for both situations, we analyze two different scenarios. First, we analyze markets in which firms can freely choose their headline prices. Second, we analyze markets in which firms are restricted to charge non-negative headline prices.

Our first main result demonstrates that there exists a fundamental trade-off between consumer protection and efficiency: Independent of whether the headline price is restricted or not, the strongest possible consumer protection policy maximizes consumer surplus but is never welfare-optimal. Consumer protection is maximized with the strongest possible hard intervention, i.e., a ban of partitioned pricing, or the strongest possible soft intervention, i.e., making additional prices fully transparent. In both cases, firms' pricing strategies cannot be deceptive, and consumers can make fully informed choices. Due to the imperfect competition between firms

arising from imperfect substitutability of products, it directly follows that equilibrium demand is inefficiently low.

In contrast, allowing firms to engage in partitioned pricing and leaving consumers with some positive degree of naivety renders consumer protection imperfect. In equilibrium, consumers perceive total prices to be lower than they actually are and, therefore, demand higher quantities. Via this channel, imperfect consumer protection boosts demand and brings equilibrium quantities closer to efficiency. Accordingly, maximizing welfare calls for imperfect consumer protection.

Although the trade-off between consumer protection and efficiency arises with hard and soft interventions and the presence and the absence of a price floor on the headline price, the strength of the welfare-maximizing intervention and the resulting welfare level strongly depend on the scenario and the type of intervention considered. First, regarding soft interventions, it holds that the welfare-maximizing degree of price transparency induces efficient demand and, hence, achieves first-best welfare only if the regulation of the additional price is sufficiently weak. However, under a rather strict additional price regulation, firms need to set comparably high headline prices to attain sufficient revenues per unit sold. Accordingly, even if prices are fully intransparent and firms can offer negative headline prices, firms do not find it optimal to offer such low prices. Hence, irrespective of whether the headline price is restricted, it is impossible to reach efficient demand if the additional price regulation is strict.

Whether it is possible to induce efficient demand with a hard intervention, i.e., a regulation of additional prices, may crucially depend on the scenario analyzed. If headline prices are unrestricted, the welfare-maximizing upper bound on the additional price induces first-best. Intuitively, the less regulated the additional price is, the stronger firms can trick consumers into overbuying by increasing the additional price and, at the same time, decreasing the headline price. It is then always possible to relax the regulation of additional prices just as much as necessary to increase demand up to efficiency. Furthermore, our results reveal that not to regulate partitioned pricing is the worst possible option: every bound on the additional price leads to an improvement in both welfare and consumer surplus.

If the headline price is restricted to be non-negative, it is not always possible to counteract an increase in the additional price with a decrease in the headline price as the price floor starts to be binding. If consumers are sufficiently sophisticated and, hence, per se do not strongly fall victim to overbuying, the binding price floor prevents equilibrium demand from strictly increasing with the upper bound on the additional price. Accordingly, even with unregulated partitioned pricing, equilibrium demand is inefficiently low. We also show that compared to an unregulated additional price, capping the additional price at the firms' marginal costs is welfare neutral but leads to a higher consumer surplus. This result is robust to all parameter constellations of the model. In the case of doubt, this policy is a safe course of action.

Moreover, our results allow us to shed light on the interplay between hard and soft interventions. It turns out that the strategic dependence of the two types of interventions is strongly affected by market characteristics, i.e., whether a floor for the headline price exists or not. With unrestricted headline prices, consumer protection via (i) increased price transparency and (ii) stronger price regulations are strategic substitutes with respect to welfare: an increase in the degree of price transparency calls for weaker price regulations. The negative effect of an increase in price transparency on demand has to be offset by a less strict regulation of additional prices, allowing firms to maintain high demand. However, in markets with restricted headline prices, the two types of interventions may also constitute strategic complements. If the price floor for the headline price is binding, equilibrium demand decreases in the upper bound on the additional price: contrary to the case of a non-binding price floor, consumers correctly assess that total prices are increasing in the upper bound. Accordingly, the demand-decreasing effect of an increase in price transparency needs to be offset by stronger regulations of partitioned pricing.

We contribute to a strand of the literature that discusses the effects of consumers' misperceptions of prices or products on market outcomes (for instance, Gabaix and Laibson, 2006; Chetty, 2009; Heidhues, Kőszegi, and Murooka, 2016b). In particular, this paper relates to those studies that analyze the effects of hard interventions (e.g., Heidhues, Kőszegi, and Murooka, 2016a; Heidhues, Johnen, and Kőszegi, 2020), soft interventions (e.g., Glaeser and Ujhelyi, 2010; de Meza and

Reyniers, 2012; Kosfeld and Schüwer, 2016), or both (Armstrong and Vickers, 2012) on consumer surplus or welfare.

Heidhues, Johnen, and Kőszegi (2020) compare the effects of two hard interventions, regulating total prices and regulating additional prices in which consumers can choose between browsing and studying. In this model, consumers have limited attention such that they can only examine a few products in detail, including their additional prices, and have to browse other products superficially. If additional prices are constrained, consumers can engage in more browsing and, hence, compare more options. However, when total prices are regulated, the consumer is tempted to search less, which might leave him worse off. In our model, we adopt the notion that consumers are naive, as in Heidhues, Kőszegi, and Murooka (2016b). More precisely, even if they observe all prices, they still underestimate the total price as long as prices are partitioned.

In models of additional prices and hidden fees, more similar to our approach, Armstrong and Vickers (2012) and Kosfeld and Schüwer (2016) find that transparency-enhancing innovations in the form of consumer education might decrease welfare. These studies differentiate between sophisticated and naive consumers and endow sophisticates with the possibility to avoid a product's additional price at some cost. As (i) this cost is assumed to be socially wasteful, and (ii) sophisticated consumers make use of this option in equilibrium, educating naifs to sophisticates might be welfare-decreasing. Moreover, Armstrong and Vickers (2012) show that in contrast to a transparency-enhancing intervention, regulation of additional prices is at least weakly welfare increasing. Because sophisticated consumers only exert socially wasteful effort if additional prices are sufficiently high, tight regulation of additional prices avoids socially wasteful effort and welfare increases. However, Heidhues, Kőszegi, and Murooka (2016a) show that this might lead to higher innovation incentives on new types of hidden charges, which then lowers the policy's welfare effect. While these studies focus on inefficiencies that arise due to costs that sophisticated consumers bear to avoid additional charges, welfare effects, in our model, are solely driven by inefficiently low or high demand.

Glaeser and Ujhelyi (2010) analyze a model of Cournot competition with positive profits and inefficiently low demand in equilibrium. They show that consumers' underestimation of future health costs caused by consumption is detrimental to consumer surplus but leads to a more efficient demand. Accordingly, the mechanism at work is similar to the channel through which the negative effect of price transparency on welfare emerges in our paper. However, in their model, consumers perceive prices correctly and only underestimate the cost they have to bear. Thus, they remain silent about potential hard interventions.

Most closely related to our analysis is de Meza and Reyniers (2012). They show that when consumers underestimate additional prices in a Cournot model with constant elasticity of demand, consumer surplus and, with it, welfare might decrease in transparency. Full transparency leads to an increased and, hence, more inefficient total price. However, while full transparency also leads to an inefficiently high total price in our model, their result crucially hinges on the assumption of isoelastic demand and competition in quantities. We complement their study by discussing partitioned pricing and consumer naivety in the classical framework by Singh and Vives (1984). Moreover, we discuss the effects of hard and soft interventions, their interplay, and identify a trade-off between consumer surplus and efficiency.

The paper is organized as follows. Section 2.2 introduces the model. In Section 2.3, we solve for the equilibrium and do the comparative statics analysis for the case of unrestricted headline prices (2.3.1) and when non-negative headline prices are not feasible (2.3.2). Finally, Section 2.4 concludes.

2.2 Model

We adopt the differentiated duopoly framework by Singh and Vives (1984), in which each of the two firms $i \in \{1,2\}$ produce quantity q_i of good i at constant marginal cost c. There is a continuum of consumers with mass one who derive utility from consuming quantities $q = (q_1, q_2)$ via

$$U(q) = \omega \sum_{i=1}^{2} q_i - 0.5\left(\sum_{i=1}^{2} q_i^2 + 2\gamma q_i q_j\right) \text{ for } i \in \{1,2\} \text{ and } j \neq i,$$

where q_i denotes the quantity purchased at firm i, $\omega > 0$ is a measure of product quality, and $\gamma \in (-1, 1)$ measures the degree of substitutability between the two products. The higher the value of γ, the more alike the products are.

Firms compete via prices to attract consumers. We extend the original framework by incorporating partitioned pricing and consumer naivety. First, we allow firms to partition prices, i.e., we assume that the total price of good i is given by the sum of a headline price p_i and an additional price $\hat{p}_i$. Among others, we follow Gabaix and Laibson (2006) in assuming bounded additional prices, i.e. $\hat{p}_i \leq \bar{p}$. Essentially, the assumption limits firms' ability to exploit naive consumers as they cannot be tricked into buying an infinite amount of a good. In our model, we treat $\bar{p}$ as a policy measure that can be influenced by policymakers.[8] We label the two extreme cases $\bar{p} = 0$ and $\bar{p} \to \infty$ as a ban of partitioned pricing and unregulated additional prices, respectively.

Second, consumer base their buying decision on a perceived price $p_i + \beta \cdot \hat{p}_i$ with $\beta \in [0, 1]$ instead of the actual total price $p_i + \hat{p}_i$. Note that β measures the degree of sophistication of the consumers in our model, i.e., the higher β, the closer is the consumers' perceived price to the actual price $p_i + \hat{p}_i$. Therefore, we can interpret β as a policy parameter that influences firms' ability to charge intransparent additional prices, e.g., a minimum font size of an additional mandatory fee on a price comparison website. We refer to the extreme cases $\beta = 1$ and $\beta = 0$ as fully transparent and fully intransparent prices, respectively.

Since consumers underestimate the total price if firms engage in partitioned pricing they maximize their *perceived* net-utility function

$$\tilde{V}\left(q, p, \hat{p}\right) = U(q) - \sum_{i=1}^{2} q_i \left(p_i + \beta \hat{p}_i\right).$$

with $p = \left(p_i, p_j\right)$, and $\hat{p} = \left(\hat{p}_i, \hat{p}_j\right)$, when making their buying decisions. Consumer behavior is defined by the first-order conditions[9] given by

8. Although we treat $\bar{p}$ as a policy variable, such an upper bound could also be imposed by other players in the market. For example, car loan companies allow car dealers to mark-up loans up to some percentage points but not above (Grunewald, Lanning, Low, and Salz, 2020).

9. Since $\frac{\partial^2 \tilde{V}\left(q,p,\hat{p}\right)}{\partial q_i^2} = \frac{\partial^2 \tilde{V}\left(q,p,\hat{p}\right)}{\partial q_j^2} = -1$, $\frac{\partial^2 \tilde{V}\left(q,p,\hat{p}\right)}{\partial q_j \partial q_i} = -\gamma$, and $det\,(H) = 1 - \gamma^2 > 0$, where $det\,(H)$ denotes the Hessian matrix, the consumers' perceived utility is a strictly concave function.

$$\frac{\partial \tilde{V}\left(q,p,\hat{p}\right)}{\partial q_i} = \omega - q_i - \gamma q_j - p_i - \beta \hat{p}_i \stackrel{!}{=} 0 \qquad \forall\, i \in \{1,2\}.$$

The resulting symmetric demand functions are given by

$$q_i\left(p,\hat{p}\right) = \frac{1}{1-\gamma^2}\left[(1-\gamma)\omega - (p_i + \beta\hat{p}_i) + \gamma(p_j + \beta\hat{p}_j)\right] \quad i \neq j, i,j \in \{1,2\}$$

Firm profits are, therefore, given by

$$\pi_i(p,\hat{p}) = q_i\left(p,\hat{p}\right) \cdot \left(p_i + \hat{p}_i - c\right) \qquad \forall\, i \in \{1,2\}. \tag{2.1}$$

We follow the usual definition of producer surplus as the sum of profits, i.e.

$$\mathscr{PS} = \sum_{i=1}^{2} \pi_i(p,\hat{p}). \tag{2.2}$$

While the definition of producer surplus is standard in our model, we need to take a stance on how consumer surplus is measured because they are naive. Consumers base their buying decision on the perceived total price. As they end up, however, paying the actual total price, we follow Glaeser and Ujhelyi (2010) and measure consumer surplus as the experienced net-utility, i.e.,

$$\mathscr{CS} = U(q) - \sum_{i=1}^{2} q_i\left(p_i + \hat{p}_i\right).$$

Welfare is defined as the sum of producer and consumer surplus such that

$$\mathscr{W} = \omega \sum_{i=1}^{2} q_i - 0.5\left(\sum_{i=1}^{2}(q_i)^2 + 2\gamma q_i q_j\right) - \sum_{i=1}^{2} q_i \cdot c.$$

Maximizing welfare[10] over quantities yields the first-best quantities, given by $q_i^{FB} = \frac{\omega - c}{1+\gamma}$, $\forall i \in \{1,2\}$, which will serve as a benchmark in the following analysis.

As argued in the introduction, authorities have mainly engaged in two policy measures to protect consumers from being deceived: increasing price transparency (increasing β) and decreasing additional prices (decreasing $\bar{p}$). These policies aim at reducing the wedge between the perceived and the actual prices and, thereby,

10. Note that welfare is a strictly concave function since $\frac{\partial^2 \mathscr{W}}{\partial q_i^2} = \frac{\partial^2 \mathscr{W}}{\partial q_j^2} = -1$, $\frac{\partial^2 \mathscr{W}}{\partial q_j \partial q_i} = -\gamma$, and $det\,(H) = 1-\gamma^2 > 0$, where $det\,(H)$ denotes the determinant of the Hessian matrix.

allow consumers to make more informed choices. In line with the rationale behind these policies, we think of consumer protection as a measure inversely related to the differences between the perceived and the actual prices, $(1-\beta)\hat{p}_i$. [11] Intuitively, the higher the level of consumer protection, the closer are the perceived prices to the actual prices. Full consumer protection refers to the case when the perceived prices equal the actual prices, which occurs under full price transparency ($\beta = 1$) or a ban of partitioned pricing ($\bar{p} = 0$).

2.3 Policy Analysis

To analyze the welfare consequences of changing the regulatory requirements for additional prices and price transparency, we first analyze markets in which firms can partition prices and face no restriction on headline prices. We derive optimal policies for consumers, producers, and welfare in these types of markets. Afterward, we consider markets where headline prices are restricted to be non-negative and perform the same analysis. Our results hinge on crucial similarities and differences between consumer optimal and welfare optimal policies in both types of markets.

2.3.1 Unrestricted Headline Prices

In markets in which firms can freely choose their headline prices, firms only take the restriction on the additional price into account when choosing their prices. Therefore, anticipating the demand of its consumers, firm i chooses its prices to solve

$$\max_{p_i,\hat{p}_i} \pi_i(p,\hat{p}) \;=\; q_i\big(p,\hat{p}\big)\cdot\big(p_i+\hat{p}_i-c\big) \;\; s.t. \;\; \hat{p}_i \;\leq\; \bar{p}.$$

Since consumers are less sensitive to an increase in the additional price than to a corresponding decrease in the headline price, firms will always choose the highest possible additional price, i.e., in any equilibrium $\hat{p}_i^* = \bar{p}$. To see this, note that for any combination of prices $(p_i,\hat{p}_i)$ of firm i with $\hat{p}_i < \bar{p}$, there exists a feasible combination of prices $(p_i^{'},\hat{p}_i^{'})$ with $\bar{p} \geq \hat{p}_i^{'} > \hat{p}_i$ and $p_i^{'} = p_i + \hat{p}_i - \hat{p}_i^{'}$ that leads to a strictly

11. More formally, we define consumer protection as a differentiable function $\mathscr{P} : [0,\bar{p}]^2 \to \mathbb{R}$, with strictly negative partial derivatives and the arguments being the difference between the actual and perceived total prices, $(1-\beta)\hat{p}_i \;\forall\; i$.

higher profit: revenue per unit sold remains constant but demand strictly increases. Consequently, firm i's maximization problem reduces to

$$\max_{p_i} q_i\left(p, \hat{p}\right) \cdot \left(p_i + \bar{p} - c\right).$$

The best-response function of firm $i \in \{1, 2\}$ is then given by

$$p_i = \frac{\left(1-\gamma\right)\omega}{2} + \frac{\gamma}{2} \cdot p_j + \frac{c - \left[1 + \left(1-\gamma\right)\beta\right]\bar{p}}{2}$$

for $i \in \{1, 2\}$ and $j \neq i$. Calculating the equilibrium prices leads to

$$p_i^* = p_j^* = \frac{\left(1-\gamma\right)\omega + c - \left[1 + \left(1-\gamma\right)\beta\right]\bar{p}}{2-\gamma}, \tag{2.3}$$

which yields the following proposition:

Proposition 2.1. *The unique equilibrium is symmetric with* $\hat{p}^* = \bar{p}$*,*

$$p^* = \frac{\left(1-\gamma\right)\left(\omega - \beta\bar{p}\right) + c - \bar{p}}{2-\gamma}, \text{ and } q^* = \frac{\omega - c + \left(1-\beta\right)\bar{p}}{\left(1+\gamma\right)\left(2-\gamma\right)}. \tag{2.4}$$

Proposition 2.1 implies that an increase in the upper bound on the additional price $\bar{p}$ leads to lower headline prices p^*, higher additional prices $\hat{p}^*$, and a higher quantity q^*. As we have already shown, firms always charge the highest possible additional price. Accordingly, relaxing the upper bound, $\bar{p}$, leads to higher additional prices, which puts higher competitive pressure on the headline prices. Intuitively, the higher the additional price, the more profit can be extracted from naive consumers. Consequently, attracting consumers with low headline prices becomes more valuable, and headline prices decrease. However, as we consider differentiated products, the competitive pressure is not strong enough to fully compete away these additional profits. Accordingly, the decrease in the headline prices is smaller than the increase in the additional prices such that an increase in $\bar{p}$ leads to higher total prices $p^* + \hat{p}^*$. Consumers correctly take into account the decrease in the headline price but undervalue the increase in the additional price. It turns out that the decrease in the headline price outweighs the consumers' perceived increase in the additional price such that the perceived total price $p^* + \beta\hat{p}^*$ decreases. Hence, consumers demand

higher quantities, which explains the counter-intuitive result that equilibrium quantities and total prices increase as a response to an increase in $\bar{p}$.

A higher degree of transparency, i.e., an increase in β, does not affect the additional price but leads to lower headline prices and a lower quantity. Clearly, the additional price remains constant as it depends only on the upper bound. Higher transparency makes consumers more sensitive to a change in the additional prices. Hence, competitive forces become fiercer and lead to lower headline prices. As an increase in β only leads to a lower headline price, the total price decreases such that one would expect demand to increase. Due to higher transparency of prices, however, consumers take the additional price more strongly into account and, via this channel, perceive the good to become more expensive. This demand-decreasing effect dominates the demand-increasing effect of a lower headline price such that the perceived total price increases. Although an increase in transparency leads to lower actual prices, equilibrium quantities decrease as well.

We will continue by analyzing the effects of (i) a hard intervention via tighter regulation of additional prices (i.e., a decrease in $\bar{p}$) and (ii) a transparency-enhancing soft intervention (i.e., an increase in β). The effects of an increase in the degree of transparency given by β and of a decrease in the upper bound, $\bar{p}$ can be summarized as an increase in consumer protection: although the total price decreases, consumers perceive the good to be more expensive such that they demand lower quantities in equilibrium. Thus, both effects lead to a decrease in the wedge between actual and perceived prices, $(1-\beta)\bar{p}$. The following proposition summarizes its effects on consumer surplus, producer surplus, and welfare. We denote the welfare-maximizing degree of price transparency by $\beta^{SB}\left(\bar{p}\right)$ and, correspondingly, the welfare-maximizing upper bound on the additional price by $\bar{p}^{SB}\left(\beta\right)$.[12]

Proposition 2.2. *Full consumer protection maximizes consumer surplus, while no consumer protection maximizes producer surplus. Full consumer protection is not welfare-maximizing. For every $\beta < 1$, $\bar{p}^{SB}\left(\beta\right)$ induces first-best welfare. First-best can be achieved via $\beta^{SB}\left(\bar{p}\right) \in (0,1)$ if and only if $\bar{p} > \left(1-\gamma\right)(\omega - c)$. Otherwise, $\beta^{SB}\left(\bar{p}\right) = 0$.*

12. We relegate the proofs of all following propositions and lemmas to the appendix.

Confirming intuition, consumer surplus is globally increasing in the degree of consumer protection. Decreasing the wedge between actual and perceived additional prices by either decreasing $\bar{p}$ or increasing β allows consumers to make more informed choices. Therefore, they benefit from stronger consumer protection. Likewise, producer surplus is globally decreasing in the degree of consumer protection. Consumer protection mitigates the extent to which consumers underestimate the additional price, limiting the firms' possibility to exploit consumers profitably, and, hence, decreases their profits.

Importantly, Proposition 2.2 demonstrates a fundamental trade-off between consumer protection and efficiency. Protecting consumers completely from being exploited – either by banning additional prices or by eliminating price intransparency – maximizes consumer surplus. However, at the same time, such a policy renders welfare inefficiently low. In the absence of possible consumer exploitation, imperfect substitutability between the firms' products implies imperfect competition, which leads to inefficiently high equilibrium prices and eventually to inefficiently low demand. A marginal increase in $\bar{p}$ or a marginal decrease in β renders consumer protection imperfect and allows firms to take advantage of consumer naivety. Although the actual total price increases, consumers mistakenly perceive the product to become less expensive and demand higher quantities. Via this channel, imperfect consumer protection boosts demand and is strictly welfare-increasing as it mitigates the inefficiency arising from imperfect competition.

Although the effects of changes in the upper bound on the additional price and changes in the degree of price transparency are similar, an important distinction between the two policies has to be made: Whereas for every β the upper bound $\bar{p}$ can be adjusted to induce first-best welfare, the reverse is not true. For any degree of price transparency, $\beta \in [0, 1)$, equilibrium demand is increasing in $\bar{p}$. Firms react to a decrease in consumer protection by adapting equilibrium prices such that consumers mistakenly perceive the total price to decrease and demand higher quantities. Hence, relaxing the upper bound on the additional price leads to an increase in equilibrium quantities. In particular, it is always possible to select $\bar{p}$ large enough to induce exactly the demand that leads to first-best welfare.

However, the same logic does not apply regarding the impact of a change in price transparency for a given upper bound on the additional price. Even if consumers completely neglect the additional price, i.e., $\beta = 0$, extractable profits per unit via the additional price are bounded by $\bar{p}$. Accordingly, in equilibrium, firms are not willing to offer arbitrarily low headline prices. The perceived total price is bounded from below, and the consumers' demand is, therefore, bounded from above. If $\bar{p}$ is sufficiently large, the upper bound on demand exceeds efficient demand, and the welfare-maximizing degree of price transparency $\beta^{SB}\left(\bar{p}\right) \in (0,1)$ induces first-best. However, with a strongly regulated additional price it might not be possible to adjust the degree of price transparency to achieve the first-best welfare. In equilibrium, firms can charge only low additional prices such that they need to charge comparably high headline prices. Accordingly, even with minimal price transparency, the perceived total price is so high that equilibrium demand is inefficiently low. The trade-off between consumer surplus and efficiency then globally holds as every policy designed to increase price transparency leads to an increase in consumer surplus but is detrimental to welfare. Complementing Proposition 2.2, the following lemma characterizes the welfare-maximizing regulations.

Lemma 2.1. *If $\beta < 1$, then the equilibrium welfare-maximizing upper bound on the additional price is given by $\bar{p}^{SB}\left(\beta\right) = \frac{(1-\gamma)(\omega-c)}{1-\beta}$. Similarly, if $\bar{p} > 0$, then $\beta^{SB}\left(\bar{p}\right) = \max\left\{0, 1 - \frac{(1-\gamma)(\omega-c)}{\bar{p}}\right\}$ denotes the welfare-maximizing degree of price transparency.*

The equation for $\bar{p}^{SB}\left(\beta\right)$ reveals that if prices are rather intransparent, i.e., β is low, only moderate additional prices should be allowed to induce first-best: consumers perceive total prices to be low and demand relatively high quantities per se such that a high upper bound on the additional price would result in excessive demand. Ceteris paribus, the more transparent prices are, the lower the quantities consumers demand. Accordingly, the more transparent prices are, the less regulated additional prices should be because an increase in $\bar{p}$ leads to lower perceived total prices, which offsets the decrease in demand caused by higher price transparency. Therefore, $\bar{p}^{SB}\left(\beta\right)$ is increasing in the degree of price transparency. In analogy, the welfare-maximizing degree of price transparency $\beta^{SB}\left(\bar{p}\right)$ is increasing in $\bar{p}$. The re-

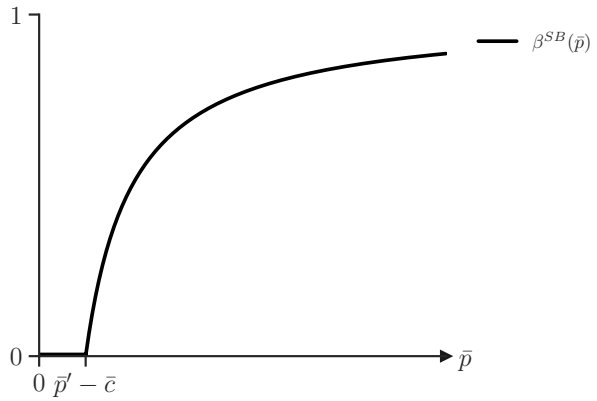


Figure 2.1. The equilibrium welfare-maximizing degree of price transparency (β^{SB}) as a function of the upper bound on the additional price ($\bar{p}$) in markets with unrestricted headline prices.

lationship is strict for $\bar{p}$ being sufficiently large. However, if the upper bound on the additional price is so low that first-best welfare is infeasible, the welfare-maximizing degree of transparency is zero and, hence, independent of $\bar{p}$.

Interestingly, our model predicts that in markets with fiercer competition via a higher degree of substitutability between the products (higher γ), ensuring welfare-maximizing demand requires stricter regulations regarding the additional price or price transparency. Fiercer competition leads to lower prices, higher quantities, and, thus, higher welfare even in the absence of strict regulations. Consequently, equilibrium quantities are close to being efficient, and strong consumer protection policies are needed to prevent excessive demand. When the degree of substitutability between the products and the degree of competition is low, ensuring efficient demand requires some leeway for firms regarding deceptive pricing practices and calls for a weaker consumer protection policy to ensure efficiency. An immediate implication of these observations is that the trade-off between consumer surplus and welfare is less severe in more competitive markets. In these markets, agencies should always regulate additional prices strongly independent of which standard they apply.

2.3.2 Non-Negative Headline Prices

A crucial feature of the equilibrium described in Proposition 2.1 is that a higher degree of price transparency or a weaker upper bound on the additional price leads to a lower headline price. Accordingly, with very weak or non-existing regulations of the additional prices, the equilibrium headline price can be arbitrarily low. While this might be possible in some markets, in others, negative headline prices might prove infeasible. For example, in the case in which firms' pricing strategy involves a base product and an unavoidable add-on product, Heidhues, Kőszegi, and Murooka (2016a,b) argue that arbitrage traders, i.e., consumers that only buy the base good, can prevent firms from charging negative headline prices. They also argue that customers might become suspicious and abstain from buying a product when confronted with overly low headline prices. Moreover, an effective price floor may also arise as a consequence of legal restrictions, e.g., a ban of below-cost pricing, anti-dumping duties, or anti-dilution clauses for mutual funds.[13]

To account for these restrictions, we will extend the model analyzed in section 2.3.1 and follow Armstrong and Vickers (2012), Grubb (2014), and Heidhues, Kőszegi, and Murooka (2016a,b) by assuming that headline prices have to be non-negative, i.e., $p_i \geq 0$ for all i. As the firms' ability to sell higher quantities with higher total prices crucially depends on the possibility of decreasing headline prices while increasing additional prices, one might be tempted to conclude that imposing a floor on the headline price changes our results from the previous sections. While the results in the following section will prove the robustness of our main insights, as the fundamental trade-off between consumer surplus and welfare prevails, a price floor has important implications for welfare-optimal policies. For example, the price floor's existence sometimes renders the implementation of first-best welfare with a suitable upper bound on the additional price infeasible and may drastically affect the design of a regulatory intervention that maximizes equilibrium welfare. Therefore, it proves necessary for regulators to distinguish between markets in which firms

13. See, for instance, https://cutt.ly/ChLg1Pi on a ban of below-cost pricing in the United States.

face a lower price bound for headline prices and markets in which headline prices are essentially unrestricted when designing welfare-optimal policies.

Deriving equilibrium prices is more subtle in this variant of the model because it is not immediately clear that firms choose to price at the same bounds of the prices. However, as in the previous section, firms never find it optimal to set two interior prices. Hence, the best-response function of firm i either involves the highest possible additional price, the lowest possible headline price, or both. The following proposition characterizes the equilibrium with a lower bound for the headline price.

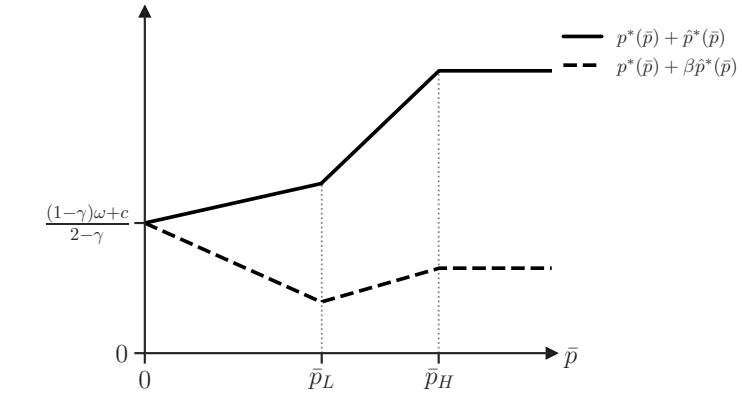
Proposition 2.3. *Let* $\bar{p}_L = \frac{(1-\gamma)\omega+c}{1+(1-\gamma)\beta}$ *and* $\bar{p}_H = \frac{(1-\gamma)\omega+\beta c}{\beta(2-\gamma)}$ *for* $\beta > 0$*. Then the unique equilibrium is symmetric and given by*

(i) $p^* = \frac{(1-\gamma)\omega+c-(1+(1-\gamma)\beta)\bar{p}}{2-\gamma}$, $\hat{p}^* = \bar{p}$, *and* $q^* = \frac{\omega-c+(1-\beta)\bar{p}}{(1+\gamma)(2-\gamma)}$ *if* $\bar{p} < \bar{p}_L$,

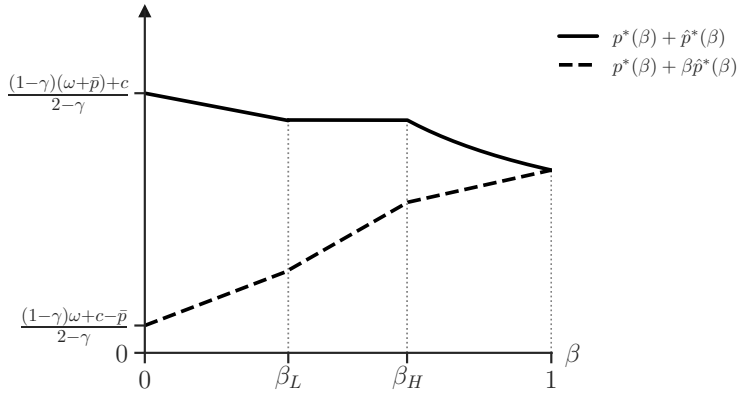
(ii) $p^* = 0$, $\hat{p}^* = \bar{p}$, *and* $q^* = \frac{\omega-\beta\bar{p}}{1+\gamma}$ *if* $\bar{p} \in [\bar{p}_L, \bar{p}_H]$,

(iii) and $p^* = 0$, $\hat{p}^* = \frac{(1-\gamma)\omega+\beta c}{(2-\gamma)\beta}$, *and* $q^* = \frac{\omega-\beta c}{(1+\gamma)(2-\gamma)}$ *if* $\beta > 0$ *and* $\bar{p}_H < \bar{p}$.

The proposition reveals that we have to distinguish between three different equilibrium outcomes depending on the upper price bound. If the upper bound on the additional price is sufficiently low, the equilibrium is identical to the case with unrestricted headline prices: the additional price is at its upper bound, and the headline price is positive and decreasing in $\bar{p}$. However, if the upper bound exceeds $\bar{p}_L$, the optimal unrestricted headline price is negative such that the lower bound on the headline price starts to be binding. It is then optimal to charge both prices at their respective bounds. In analogy to the previous analysis, further relaxing the upper bound on the additional price leads to higher additional prices in equilibrium. In this case, contrary to the equilibrium described in Proposition 2.2, an increase in the additional price is no longer accompanied by a decrease in the headline price. Consequently, the actual total prices and the perceived total prices increase, which implies that equilibrium quantities decrease. The less-regulated the additional prices are, the lower quantities consumers then demand. If the additional price regulation is sufficiently weak, i.e., $\bar{p} > \bar{p}_H$, the additional revenues per unit sold associated with a higher additional price can no longer compensate for the negative effect on



(a) $\omega = 2$, $\bar{c} = 1$, $\gamma = 0.5$, and $\beta = 0.3$



(b) $\omega = 2$, $\bar{c} = 1$, $\gamma = 0.5$, and $\bar{p} = 1.7$

Figure 2.2. The actual total equilibrium prices ($p^* + \hat{p}^*$) and the perceived total equilibrium prices ($p^* + \beta\hat{p}^*$) as functions of the upper bound on the additional price (2.2a) and price transparency (2.2b) in markets with restricted headline prices.

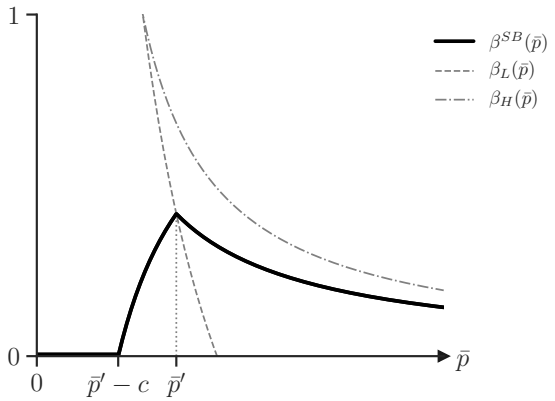
profits via lower demand. Firms prefer not to charge the maximum additional price anymore. The equilibrium is then not affected by further relaxing the upper bound on the additional price. The following proposition summarizes the implications for the effects of consumer protection on consumer surplus, producer surplus, and welfare.

Proposition 2.4. *Full consumer protection maximizes consumer surplus, while no consumer protection maximizes producer surplus. Full consumer protection is never welfare-maximizing.* $\bar{p}^{SB}\left(\beta\right) > 0$ *induces first-best if and only if* $\beta \leq \frac{c}{(1-\gamma)\omega+\gamma c}$. $\beta^{SB}\left(\bar{p}\right) \in (0,1)$ *induces first-best if and only if* $\bar{p} > \left(1-\gamma\right)(\omega - c)$.

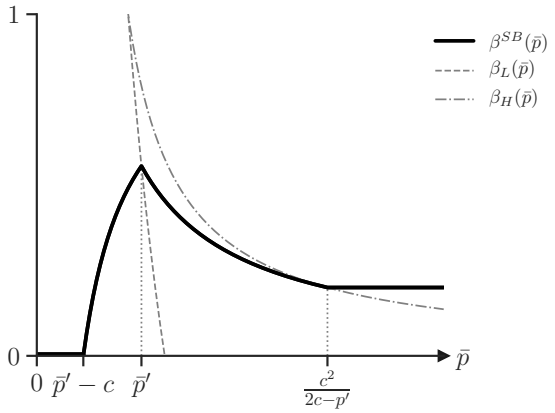
Imposing a lower bound on the headline price does not change our results from the previous section regarding consumer protection's impact on consumer and producer surplus. Protecting consumers via increased price transparency or stronger regulation of additional prices reduces the extent to which they misperceive prices and allows them to demand quantities closer to actual utility-maximizing demand. Clearly, firms suffer from consumer protection as it narrows their ability to trick consumers into overbuying profitably.

Even with a regulated headline price, it holds that full consumer protection is never welfare-maximizing. In contrast to the set-up with unrestricted headline prices, however, there does not always exist an upper bound on the additional price that leads to first-best welfare. The crucial effect of a lower bound on the headline price is that it prevents equilibrium demand from increasing monotonically with $\bar{p}$. If $\bar{p} \geq \bar{p}_L$, equilibrium demand is (weakly) decreasing in $\bar{p}$ as the headline price is at its lower bound. Accordingly, equilibrium demand is maximized with $\bar{p} = \bar{p}_L$, the lowest upper bound on the additional price that leads to a zero headline price. If imposing this upper bound leads to inefficiently low demand, achieving first-best welfare is not possible. Perversely, this is true if and only if prices are sufficiently transparent and consumers' demand is determined by rather sophisticated decision-making. This result demonstrates an essential interaction between the two policies analyzed: the higher the degree of price transparency, the less likely it is that additional prices can be regulated in a way that induces first-best welfare.

A closer look at the welfare-optimal policies facilitates our understanding of the interaction between the two policy variables. It turns out that whether a suitably-chosen degree of price transparency can achieve first-best welfare is not affected by a lower bound on the headline price. It is still valid that first-best welfare can only be achieved if the upper bound on the additional price is sufficiently large. Otherwise, demand is inefficiently low even with fully intransparent prices. In that case, the already described global trade-off between consumer surplus and welfare arises. Although the restriction on the headline price does not affect *if* it is possible to achieve first-best, it strongly affects *how* it is achieved as the critical value of β



(a) $\bar{p}' > 2c$ with $\omega = 3$, $\bar{c} = 1$, and $\gamma = 0.3$



(b) $\bar{p}' < 2c$ with $\omega = 3$, $\bar{c} = 1$, and $\gamma = 0.6$

Figure 2.3. The equilibrium welfare-maximizing degree of price transparency (β^{SB}) as a function of the upper bound on the additional price ($\bar{p}$) for $\bar{p}' > 2c$ (2.3a) and $\bar{p}' < 2c$ (2.3b) in markets with restricted headline prices.

that induces first-best, changes and differently depends on the upper bound on the additional price. The next lemma summarizes these insights.

Lemma 2.2. *Let* $\bar{p}' \equiv \left(1-\gamma\right)\omega + \gamma c$.

(i) $\beta^{SB}\left(\bar{p}\right)$ *is unique for all* $\bar{p} > 0$.

(ii) If $0 < \bar{p} \leq \bar{p}'$ *then* $\beta^{SB}\left(\bar{p}\right)$*, increases in* $\bar{p}$*. The relation is strict if* $\bar{p}' - c < \bar{p}$.

(iii) If $\bar{p} > \bar{p}'$, $\beta^{SB}\left(\bar{p}\right)$ *decreases in* $\bar{p}$*. The relation is strict unless* $\bar{p}' < 2c$ *and* $\bar{p} > \frac{c^2}{2c-\bar{p}'}$.

As for any $\bar{p} > 0$ the equilibrium quantities are monotonically decreasing in β and does not affect the first-best quantity the equilibrium welfare-maximizing de-

gree of price transparency is unique. The two panels of Figure 2.3 illustrate how $\beta^{SB}\left(\bar{p}\right)$ depends on $\bar{p}$. The left panel depicts $\beta^{SB}(\bar{p})$ for $\bar{p}' > 2c$, whereas the right panel depicts it for the opposite case. The dotted lines $\beta_L(\bar{p})$ and $\beta_H(\bar{p})$ delineate the different cases described in Proposition 2.3. Firms play the equilibrium (i) described in Proposition 2.3 if $\beta < \beta_L(\bar{p})$, equilibrium (ii) if $\beta_L(\bar{p}) \leq \beta \leq \beta_H(\bar{p})$, and equilibrium (iii) if $\beta > \beta_H(\bar{p})$.

First, if regulation of the additional price is sufficiently strong, i.e., $\bar{p} \leq \bar{p}'$, the results are identical to the case with unrestricted headline price. As $\beta^{SB}(\bar{p})$ is below $\beta_L(\bar{p})$, equilibrium prices are identical to the case with unrestricted headline prices. With a very low bound on the additional price, $\bar{p} \leq \bar{p}' - c$, equilibrium demand is inefficiently low irrespective of the exact degree of price transparency. Accordingly, welfare is maximized with fully intransparent prices, i.e., $\beta^{SB}\left(\bar{p}\right) = 0$. With a moderate bound on the additional price, $\bar{p}' - c < \bar{p} \leq \bar{p}'$, it is possible to induce first-best welfare with $\beta^{SB} \in (0,1)$. If the degree of transparency is set so as to maximize welfare, the perceived total equilibrium price is decreasing in $\bar{p}$. To offset increase in demand associated with an increase in $\bar{p}$, transparency needs to be increased to maintain first-best welfare. Accordingly, $\beta^{SB}\left(\bar{p}\right)$ is strictly increasing in $\bar{p}$.

However, contrary to the case with unrestricted headline prices the welfare-maximizing degree of transparency is not monotonically increasing in $\bar{p}$. If $\bar{p} > \bar{p}'$, then $\beta^{SB}(\bar{p}) > \beta_L(\bar{p})$. Under the welfare-maximizing degree of price transparency, firms then choose a headline price at the lower bound in equilibrium. Consumers correctly anticipate that the total equilibrium price increases with $\bar{p}$. Then, equilibrium demand is decreasing in the upper bound, and price transparency needs to be reduced to maintain the demand that induces first-best welfare. The welfare-maximizing degree of price transparency, β^{SB}, is then decreasing in $\bar{p}$.

It remains to analyze the limit case, i.e., how β^{SB} evolves if $\bar{p} \to \infty$. One might be tempted to conclude that there needs to exist a critical value of $\bar{p}$ from which on β^{SB} is constant and does not change with a further relaxation of the additional prices' regulation because the third case of Proposition 2.3 is reached. This scenario is depicted in the right panel of Figure 2.3: for $\bar{p}' \leq 2c$, $\beta^{SB}(\bar{p}) > \beta_H(\bar{p})$ if $\bar{p} > \frac{c^2}{2c-\bar{p}'}$. The intuition is that with a restricted headline price, firms are not willing to offer

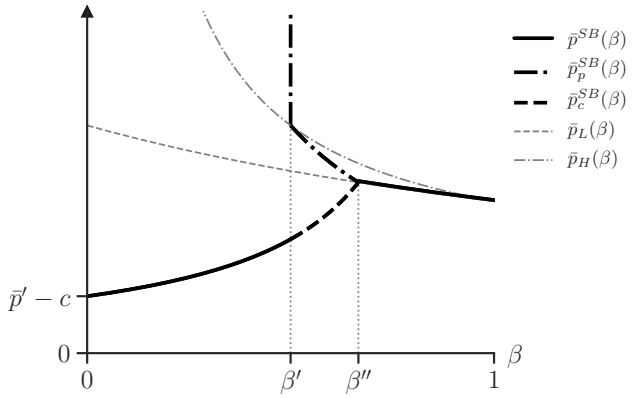
arbitrarily high additional prices. If $\bar{p}$ is sufficiently large, firms play equilibrium (iii) described in Proposition 2.3. Accordingly, they prefer to offer an interior additional price such that the regulation of additional prices is ineffective. A further increase then does not change equilibrium prices and, hence, also the welfare-maximizing degree of price transparency, β^{SB}, does not depend on $\bar{p}$.

For $\bar{p}' > 2c$, however, β^{SB} is strictly decreasing in $\bar{p}$ even for very high values of $\bar{p}$. Contrary to the previously described case, $\beta^{SB}(\bar{p}) < \beta_H(\bar{p})$ for all values of $\bar{p}$. Even with a very large bound on the additional price, the degree of price transparency that induces first best welfare is so low that firms charge maximum additional prices in this equilibrium. It then clearly holds that β^{SB} is strictly decreasing in $\bar{p}$ and approaches zero as $\bar{p} \to \infty$.

The crucial difference is that only if $\bar{p}' < 2c$, the first-best quantity is low enough such that it can be reached in equilibrium in case (iii) of Proposition 2.3. Note that for any given $\bar{p}$ the equilibrium quantity is strictly decreasing in β, because higher price transparency makes consumers more aware of additional prices and, hence, decreases their demand. As case (iii) of Proposition 2.3 is only reached if $\beta > \beta_H(\bar{p})$, the equilibrium quantity is rather low in this equilibrium. Consequently, first-best welfare can only be achieved in this case if the first-best quantity is sufficiently low. As $\bar{p}' < 2c$ is more likely to be fulfilled if the gains from trade, $\omega - c$, are low, and the degree of substitutability, γ, is high, the inequality ensures that the first-best quantities are low enough such that they are reached in case (iii) of Proposition 2.3.

Overall, it becomes evident that the welfare-maximizing degree of price transparency strongly depends on whether headline prices are restricted or not. With unrestricted headline prices and a large bound on the additional price, a regulator needs to implement (almost) fully transparent prices to induce efficient demand. This is not the case if headline prices are restricted. Instead, quite the opposite might be true as we have shown that inducing efficient demand calls for (nearly) fully intransparent additional prices if $\bar{p}' > 2c$.

We will now analyze how the welfare-optimal upper bound on the additional price, $\bar{p}^{SB}(\beta)$, depends on the degree of price transparency. Note that $\bar{p}^{SB}(\beta)$ is not



(a) $\omega = 2$, $\bar{c} = 1$, and $\gamma = 0.5$

Figure 2.4. The equilibrium welfare-maximizing upper bounds on the additional price ($\bar{p}^{SB}$, $\bar{p}_c^{SB}$, and $\bar{p}_p^{SB}$) as functions of price transparency (β) in markets with restricted headline prices.

necessarily unique. Whenever the welfare maximizing upper bound is not unique, we will, in a slight abuse of notation, refer to the set of welfare-maximizing upper bounds as $\bar{p}^{SB}\left(\beta\right)$ as well. To see that there might exist several welfare maximizing upper bounds recall that equilibrium demand is increasing in $\bar{p}$ only if $\bar{p} \leq \bar{p}_L$ and decreasing otherwise. This implies that if some $\bar{p}$ below $\bar{p}_L$ maximizes welfare, there might exist at least one additional upper bound above this threshold that leads to the same demand and, accordingly, maximizes welfare as well. The following lemma characterizes the instances under which the welfare-optimal upper bound on the additional price is unique or not as well as the effects of a change in the degree of price transparency on the welfare-optimal upper bound.

Lemma 2.3. *Let* $\beta' \equiv \left(2-\gamma\right)-\left(1-\gamma\right)\frac{\omega}{c}$ *and* $\beta'' \equiv \frac{c}{(1-\gamma)\omega+\gamma c}$.

(i) If $\beta \geq \beta''$ *then* $\bar{p}^{SB}\left(\beta\right)$ *is unique and decreasing in* β.

(ii) If $\beta \in \left(\beta', \beta''\right)$ *then* $\bar{p}^{SB}\left(\beta\right) = \left\{\bar{p}_c^{SB}\left(\beta\right), \bar{p}_p^{SB}\left(\beta\right)\right\}$ *and* $\bar{p}_c^{SB}\left(\beta\right) < \bar{p}_p^{SB}\left(\beta\right)$, *also* $\bar{p}_c^{SB}\left(\beta\right)$ *increases and* $\bar{p}_p^{SB}\left(\beta\right)$ *decreases in* β.

(iii) If $\beta = \beta'$ *then* $\bar{p}^{SB}\left(\beta\right) = \left\{\bar{p} \mid \bar{p} = \bar{p}_c^{SB}\left(\beta'\right) \vee \bar{p} \geq \bar{p}_H\left(\beta'\right)\right\}$.

(iv) If $\beta < \beta'$ *then* $\bar{p}^{SB}\left(\beta\right)$ *is unique and increasing in* β.

Figure 2.4 and Lemma 2.3 show that the optimal upper price bound is non-monotonic in β. To understand the intuition behind this result, note that by the

discussion after Proposition 2.3 the welfare-optimal upper bound on the additional price either achieves or falls short of the first-best quantity. If price transparency increases, the equilibrium quantities always decrease because consumers become more aware of the additional prices. Consequently, the welfare-optimal upper price bound has to adjust such that it counteracts the decrease in the equilibrium quantity due to an increase in price transparency. Whether an increase or a decrease of the upper bound leads to an increase in the equilibrium quantities depends on the question which of the three cases of Proposition 2.3 is at play.

Suppose first that the first-best quantities are not achievable in any of the three cases in Proposition 2.3, which is the case if $\beta > \beta''$. Intuitively, price transparency and first-best quantities are so high under these circumstances that firms cannot fool consumers into buying the first-best quantities because this would require negative headline prices in equilibrium. Therefore, the welfare optimal upper price bound is unique and given by $\bar{p}_L$ because the equilibrium quantities increase for $\bar{p} < \bar{p}_L$ and (weakly) decreases otherwise. Moreover, $\bar{p}_L$ decreases in β such that the welfare optimal upper bound decreases for $\beta > \beta''$. The reason why $\bar{p}_L$ decreases in β is that with higher price transparency, firms have to compensate consumers for higher additional prices via a stronger decrease in the headline prices such that the lower bound on the headline price is reached for lower values of $\bar{p}$.

Suppose that the first-best quantities are achievable in one of the three cases in Proposition 2.3 now, which is the case if $\beta \leq \beta''$. As argued above, the welfare-optimal upper bound is then not necessarily unique because the equilibrium quantity is a non-monotonic function of the upper bound on the additional price. When price transparency is in an intermediate range, i.e., $\beta \in \left(\beta', \beta''\right)$, then first-best quantities can be either achieved by a restrictive upper bound on the additional price, $\bar{p}_c^{SB}\left(\beta\right)$, or a less-restrictive upper bound on the additional price, $\bar{p}_p^{SB}\left(\beta\right)$. While firms charge an additional price at the respective upper bound in each of the two cases, firms charge positive headline prices in the first case and headline prices at their lower bound in the second case. In the first case, firms can still counteract an increase in the additional price by decreasing the headline price, which leads consumers to increase their demand. However, this is not feasible in the second case, where firms already

charge headline prices at their lower bound such that demand decreases in the upper bound on the additional price. Consequently, the restrictive upper price bound that achieves first-best welfare increases with an increase in price transparency, while the less-restrictive upper price bound that achieves first-best welfare decreases with an increase in price transparency for intermediate degrees of price transparency.

However, if $\beta < \beta'$, i.e. the price transparency is relatively low, then the optimal upper price bound is again unique. The reason is that price transparency is so low in this case that whenever firms would charge headline prices at their lower bound, equilibrium demand would exceed the first-best quantities. Therefore, only upper-bounds that lead to positive headline prices may induce first-best welfare. Again, the logic for the restrictive upper-bound on the additional price described in the previous paragraph applies, and the welfare-optimal bound on the additional price increases in price transparency in this case.

In the knife-edge case, when $\beta = \beta'$ Lemma 2.3 and Figure 2.4 show that there exist infinitely many upper price bounds that maximize equilibrium welfare. The reason for this result is that the equilibrium quantity is a constant function of the upper price bound on the additional price for $\bar{p} > \bar{p}_H(\beta')$. When $\beta = \beta'$ the equilibrium quantity in case (iii) of Proposition 2.3 is equal to the first-best quantity and, therefore, achieves first-best welfare. However, because firms do not find it optimal to set an additional price at the upper price bound in this case, equilibrium quantities do not depend on the upper bound on the additional price. Hence, any upper-bound on the additional price that induces case (iii) of Proposition 2.3 maximizes welfare in this case. Moreover, as the equilibrium quantity is non-monotonic in $\bar{p}$, the same logic as in the case of two welfare-maximizing upper bounds applies. Hence, there also exist a restrictive upper bound, $\bar{p}_c^{SB}(\beta')$, that maximizes equilibrium welfare.

Note that whenever there exist two equilibrium-welfare maximizing upper price bounds on the additional price the different bounds lead to the same level but to a different distribution of welfare. While $\bar{p}_c^{SB}(\beta)$ is the consumer-optimal choice among the two, $\bar{p}_p^{SB}(\beta) > \bar{p}_c^{SB}(\beta)$ is producer-optimal. Intuitively, because demand is equal in both scenarios, perceived prices must be equal as well. However, actual total prices differ in the two cases. As actual total prices only redistribute

welfare from consumers to firms, firms profit from higher actual total prices, while consumers profit from lower actual total prices. From the discussion of Proposition 2.3 we now that actual total prices are strictly increasing in the upper bound on the additional prices such that firms get a larger share of the same surplus under $\bar{p}_p^{SB}(\beta)$ compared to $\bar{p}_c^{SB}(\beta)$. This observation is important for regulators who want to promote consumer surplus but do not want to sacrifice welfare. In the case of intermediate ranges of price transparency, they should aim at regulating firms more strictly by introducing the tighter upper bound $\bar{p}_c^{SB}(\beta)$ instead of $\bar{p}_p^{SB}(\beta)$.

We complete the analysis with another observation of strong practical relevance for consumer protection and competition authorities.

Proposition 2.5. *If $\beta < 1$, compared to effectively unregulated partitioned pricing ($\bar{p} \geq \bar{p}_H$), an upper bound on the additional prices that equals the firms' marginal costs, i.e., $\bar{p} = c$, leaves welfare unaffected but yields a strictly higher consumer surplus.*

The perceived prices, which determine the consumers' purchasing decision, are decreasing in $\bar{p}$ up to $\bar{p} = \bar{p}_L$, then increase with $\bar{p}$ and converge to $\frac{(1-\gamma)\omega+\beta c}{2-\gamma}$ as the upper bound on the additional prices becomes arbitrarily large, which is identical to the perceived price in the equilibrium with $\bar{p} = c$. Hence, consumers demand exactly the same quantity if $\bar{p} = c$ or $\bar{p} \geq \bar{p}_H$. Importantly, this result is independent of the degree of substitutability between products, the gains from trade, and the degree of price transparency. Note that the actual total prices are globally increasing in $\bar{p}$. Therefore, consumers strictly benefit from a regulation on the additional prices tied to the firm's actual cost. As the effect of such a regulation is independent of the exact parameter constellations and the regulation itself is easy to formulate and implement, we consider it to be a highly relevant regulatory intervention designed to protect consumers and not to harm welfare.

2.4 Concluding Remarks

We study the impact of consumer naivety and partitioned pricing in the differentiated duopoly framework of Singh and Vives (1984) and examine the interplay between hard and soft interventions and how they affect consumer surplus and wel-

fare. Full consumer protection maximizes consumer surplus but is never welfare-maximizing. This demonstrates a trade-off between consumer protection and efficiency, which is essential for evaluating regulatory interventions to increase transparency or regulate pricing strategies. We show that this trade-off's strength depends on the gains from trade and the degree of substitutability between the two goods.

However, our results show that any regulation on the additional price increases welfare and consumer surplus, compared to no regulation in the case of unrestricted headline prices. If headline prices are restricted to be non-negative, capping additional prices at the firms' marginal costs renders welfare unaffected but makes consumers better off. Moreover, we elucidate the interplay between the welfare-optimal degree of price transparency and additional price regulation. For unregulated headline prices, these policies are substitutes. If prices are relatively transparent, additional price regulation should be sufficiently loose to reach the first-best welfare level. On the contrary, when headline prices cannot be negative, whether these policies are substitutes or complements depends on whether the bound on the headline price binds in equilibrium or not.

Given our results, a natural question for policymakers arises: How should the trade-off between consumer surplus and welfare be resolved? The majority of antitrust agencies, with Canada as a notable exception (Heyer, 2006), have primarily focused on consumer surplus when evaluating potential interventions. However, this also implies that agencies only intervene when consumers' interests are at stake and do not intervene to increase efficiency. Scholars have argued that efficiency should play a more important role when agencies decide on suitable policy measures in a recent and ongoing debate (Wilson, 2019). We do not take a stance in this important discussion, but we think that our analysis is informative for both sides, as it identifies optimal policies from a consumer and from a welfare perspective. Moreover, whenever an agency decides to give weight to both principles, our results on policies that increase either both or increase consumer surplus and do not sacrifice welfare might be especially informative.

Incorporating partitioning of prices into the influential framework of Singh and Vives (1984) paves the way for several promising avenues of future research. For

example, how deceptive pricing strategies influence collusive behavior is an open question. A repeated version of our model could be used to analyze the stability of tacit collusion. Furthermore, as partitioned prices affect demand, it certainly influences firms' incentives to invest in innovation. This could be investigated by adding a preceding stage where firms may invest in, e.g., cost-reducing R&D. We leave these important and interesting questions for future research.

Appendix 2.A Proofs

Proof of Proposition 2.2. In equilibrium, using the first-order conditions of the firms which yield $(1-\gamma^2)q^* = \bar{p} + p^* - c$, producer surplus is given by $\mathscr{PS}^* = 2(1-\gamma^2)q^{*2}$. Since $\frac{\partial q^*}{\partial \bar{p}} > 0$ and $\frac{\partial q^*}{\partial \beta} < 0$, we get that

$$\frac{\partial \mathscr{PS}^*}{\partial \bar{p}} = 4(1-\gamma^2)q^*\frac{\partial q^*}{\partial \bar{p}} > 0 \quad \text{and} \quad \frac{\partial \mathscr{PS}^*}{\partial \beta} = 4(1-\gamma^2)q^*\frac{\partial q^*}{\partial \beta} < 0.$$

Accordingly, no consumer protection via (i) $\beta = 0$ or (ii) $\bar{p} \to \infty$ maximizes producer surplus.

In equilibrium, again using the first-order conditions of the firms and plugging in equilibrium quantities, consumer surplus is given by $\mathscr{CS}^* = 2\omega q^* - \big(3-2\gamma\big)\big(1+\gamma\big)q^{*2} - 2q^*c$. Taking the derivatives with respect to $\bar{p}$ and β yields

$$\begin{aligned}\frac{\partial \mathscr{CS}^*}{\partial \bar{p}} &= 2\frac{\partial q^*}{\partial \bar{p}}\big(\omega - \big(3-2\gamma\big)\big(1+\gamma\big)q^* - c\big) \quad \text{and}\\ \frac{\partial \mathscr{CS}^*}{\partial \beta} &= 2\frac{\partial q^*}{\partial \beta}\big(\omega - \big(3-2\gamma\big)\big(1+\gamma\big)q^* - c\big).\end{aligned}$$

Hence, $\frac{\partial \mathscr{CS}^*}{\partial \bar{p}} < 0$ and $\frac{\partial \mathscr{CS}^*}{\partial \beta} > 0$ if and only if $q^* > \frac{\omega - c}{(3-2\gamma)(1+\gamma)}$, which always holds. Thus, full consumer protection via (i) $\beta = 1$ or (ii) $\bar{p} = 0$ maximizes consumer surplus.

Plugging in the equilibrium quantities into the welfare function yields equilibrium welfare. Note that equilibrium welfare is a strictly concave function in equilibrium quantities, which is maximized whenever $q^* = \frac{\omega - c}{1+\gamma} = q^{FB}$. Comparing first-best and equilibrium quantities reveals that

$$q^* = q^{FB} \iff \bar{p} = \frac{(1-\gamma)(\omega - c)}{1-\beta} \iff \beta = 1 - \frac{(1-\gamma)(\omega - c)}{\bar{p}}.$$

The welfare-maximizing upper bound $\bar{p}^{SB} = \frac{(1-\gamma)(\omega-c)}{1-\beta}$ induces first-best and is finite for every $\beta < 1$. Regarding the welfare-maximizing degree of price transparency, β^{SB}, it holds that $1 - \frac{(1-\gamma)(\omega-c)}{\bar{p}} \geq 0 \iff \bar{p} \geq (1-\gamma)(\omega - c)$. If, however, $\bar{p} < (1-\gamma)(\omega - c)$, first-best is not feasible. It then holds that $q^* < q^{FB} \forall \beta \in [0,1]$. As welfare is concave in q^* and $\frac{\partial q^*}{\partial \beta} < 0$, welfare is maximized with $\beta = 0$. It directly follows that $\beta^{SB} = \max\left\{0, 1 - \frac{(1-\gamma)(\omega-c)}{\bar{p}}\right\}$. □

Proof of Proposition 2.3. The proof of the proposition evolves in four steps. First, we argue that in any equilibrium each firm charges at least one price at its bound. Second, we show that in any equilibrium firms will choose to charge at least one price at the same bound. Third, we derive the unique equilibrium candidates for any value of the upper bound $\bar{p}$. Lastly, we prove that these candidates indeed constitute an equilibrium.

Step 1: First, note that for every firm at least one price bound is binding in equilibrium. Suppose, to the contrary, that both constraints are slack, i.e., $p_i > 0$ and $\hat{p}_i < \bar{p}$. It is possible to decrease p_i by ϵ and increase $\hat{p}_i$ by ϵ so that both constraints remain slack. This leads to a strictly higher profit: revenue per unit sold remains constant, but demand increases strictly as consumers are less sensitive to an increase in the additional price than a corresponding decrease in the headline price. Accordingly, whenever both constraints are slack, there exists a profitable deviation. It follows that the best-response function of firm i either involves the highest possible additional price, the lowest possible headline price, or both.

Step 2: (i) Suppose firm i chooses the lowest headline price, i.e. $p_i = 0$. The corresponding optimal additional price $\hat{p}_i^*(p_j, \hat{p}_j)$ solves the maximization problem

$$\max_{\hat{p}_i} \pi_i(p, \hat{p}) \;=\; q_i\big(p, \hat{p}\big) \cdot \big(\hat{p}_i - c\big) \;\; s.t.\; \hat{p}_i \;\leq\; \bar{p}. \tag{2.A.1}$$

The first-order condition is given by

$$\frac{-\beta}{1-\gamma^2}\big(\hat{p}_i - c\big) + \frac{1}{1-\gamma^2}\big[(1-\gamma)\omega - (p_i + \beta\hat{p}_i) + \gamma(p_j + \beta\hat{p}_j)\big] \;=\; 0. \tag{2.A.2}$$

As π_i is strictly concave in $\hat{p}_i$, it follows that

$$\hat{p}_i^*(p_j, \hat{p}_j) \;=\; \min\left\{\frac{\big(1-\gamma\big)\omega}{2\beta} + \frac{\gamma}{2}\hat{p}_j + \frac{\gamma}{2\beta}p_j + \frac{c}{2}, \bar{p}\right\}. \tag{2.A.3}$$

(ii) Suppose firm i chooses the highest possible additional price, i.e. $\hat{p}_i = \bar{p}$. The corresponding optimal headline price $p_i^*(p_j, \hat{p}_j)$ solves the maximization problem

$$\max_{p_i} \pi_i(p, \hat{p}) \;=\; q_i\big(p, \hat{p}\big) \cdot \big(p_i + \bar{p} - c\big) \;\; s.t.\; p_i \;\geq\; 0 \tag{2.A.4}$$

The first-order condition is given by

$$-\frac{1}{1-\gamma^2}\left(p_i+\bar{p}-c\right)+\frac{1}{1-\gamma^2}\left[(1-\gamma)\omega-(p_i+\beta\hat{p}_i)+\gamma(p_j+\beta\hat{p}_j)\right] = 0. \tag{2.A.5}$$

As π_i is strictly concave in p_i, it follows that

$$p_i^*(p_j,\hat{p}_j) = \max\left\{0,\frac{(1-\gamma)\omega}{2}-\frac{1+\beta}{2}\bar{p}+\frac{\gamma}{2}\left(p_j+\beta\hat{p}_j\right)+\frac{c}{2}\right\}. \tag{2.A.6}$$

Lemma 2.4. *If $\bar{p}>\bar{p}_L$ then in any equilibrium either $p_i^*=p_j^*=0$ or $p_i^*;p_j^*>0$.*

Proof. Suppose to the contrary that $\bar{p}>\bar{p}_L$ and there exists an equilibrium with $p_i^*>0$ and $p_j^*=0$. This implies

$$p_i^* = \frac{(1-\gamma)\omega}{2}-\frac{1+\beta}{2}\bar{p}+\frac{\gamma}{2}\beta\hat{p}_j^*+\frac{c}{2}. \tag{2.A.7}$$

As $\hat{p}_j^*\leq\bar{p}$, $p_i^*\leq\frac{(1-\gamma)\omega}{2}-\frac{1+\beta}{2}\bar{p}+\frac{\gamma}{2}\beta\bar{p}+\frac{c}{2}$. Accordingly, $p_i^*>0$ requires $\frac{(1-\gamma)\omega}{2}-\frac{1+\beta}{2}\bar{p}+\frac{\gamma}{2}\beta\bar{p}+\frac{c}{2}>0 \Rightarrow \bar{p}<\bar{p}_L$, a contradiction. □

Lemma 2.5. *If $\bar{p}\leq\bar{p}_L$ then in any equilibrium either $\hat{p}_i^*=\hat{p}_j^*=\bar{p}$ or $\hat{p}_i^*,\hat{p}_j^*<\bar{p}$.*

Proof. Suppose to the contrary that $\bar{p}\leq\bar{p}_L$ and there exists an equilibrium with $\hat{p}_j^*=\bar{p}$ and $\hat{p}_i^*<\bar{p}$. It then needs to hold that

$$\hat{p}_i^* = \frac{(1-\gamma)\omega}{2\beta}+\frac{\gamma}{2}\bar{p}+\frac{\gamma}{2\beta}p_j^*+\frac{c}{2} < \bar{p}. \tag{2.A.8}$$

Accordingly, $\hat{p}_i^*<\bar{p}$ requires $\frac{(1-\gamma)\omega}{2\beta}+\frac{\gamma}{2}\bar{p}+\frac{c}{2}<\bar{p} \iff \bar{p}>\frac{(1-\gamma)\omega+\beta c}{\beta(2-\gamma)}$. This, however, contradicts the assumption $\bar{p}\leq\bar{p}_L$, as $\frac{(1-\gamma)\omega+\beta c}{\beta(2-\gamma)}>\frac{(1-\gamma)\omega+c}{1+\beta(1-\gamma)} \iff \omega>\beta c$, which always holds. □

Step 3: From the two lemmas and the observation that at least one bound is binding it directly follows that (i) $p_i^*=p_j^*=0$, (ii) $\hat{p}_i^*=\hat{p}_j^*=\bar{p}$, or (iii) both.

(i) Suppose $p_i^*=p_j^*=0$. Firm $i\in\{1,2\}$ solves the maximization problem in (2.A.1) given $p_j=0$, such that the best-response function of firm i is given by

$$\hat{p}_i^*(\hat{p}_j) = \min\left\{\frac{(1-\gamma)\omega}{2\beta}+\frac{\gamma}{2}\hat{p}_j+\frac{c}{2},\bar{p}\right\}. \tag{2.A.9}$$

Note that $\hat{p}_i^* = \bar{p}$ if and only if $\frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\hat{p}_j^* + \frac{\gamma}{2\beta}p_j^* + \frac{c}{2} \geq \bar{p}$. Now suppose $\hat{p}_j^* = \bar{p}$ and $\hat{p}_i^* < \bar{p}$. This implies that $\hat{p}_i^* = \frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\bar{p} + \frac{c}{2} < \bar{p}$ and $\hat{p}_j^* = \frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\hat{p}_i^* + \frac{c}{2} > \bar{p}$ such that $\frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\hat{p}_i^* + \frac{c}{2} > \frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\bar{p} + \frac{c}{2}$, which yields a contradiction. Hence, only symmetric combinations $\hat{p}_1^* = \hat{p}_2^* = \hat{p}^*$ can constitute mutual best responses. There exists exactly one such combination with $p^* = \bar{p}$ if and only if $\hat{p}_i^*(\bar{p}) = \bar{p}$, i.e., $\frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\bar{p} + \frac{c}{2} \geq \bar{p} \iff \bar{p} \leq \frac{(1-\gamma)\omega+\beta c}{(2-\gamma)\beta} = \bar{p}_H$. Otherwise, the intersection of the best-response functions is determined by $\hat{p} = \frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\hat{p} + \frac{c}{2} \iff \hat{p} = \frac{(1-\gamma)\omega+\beta c}{(2-\gamma)\beta}$.

(ii) Now suppose $\hat{p}_i^* = \hat{p}_j^* = \bar{p}$. Firm $i \in \{1,2\}$ solves the maximization problem in (2.A.4) given $\hat{p}_j = \bar{p}$, such that the best-response function of firm i is given by

$$p_i^*(p_j) \;=\; \max\left\{0, \frac{\left(1-\gamma\right)\omega}{2} - \frac{1+\beta}{2}\bar{p} + \frac{\gamma}{2}\left(p_j + \beta\bar{p}\right) + \frac{c}{2}\right\}. \tag{2.A.10}$$

Note that $p_i^* = 0$ if and only if $\frac{(1-\gamma)\omega}{2} - \frac{(1+\beta)}{2}\bar{p} + \frac{\gamma}{2}\left(p_j^* + \beta\hat{p}_j^*\right) + \frac{c}{2} \leq 0$. Now suppoes $p_j^* = 0$ and $p_i^* > 0$. This implies that $p_i^* = \frac{(1-\gamma)\omega}{2} - \frac{(1+\beta)}{2}\bar{p} + \frac{\gamma}{2}\beta\bar{p} + \frac{c}{2} > 0$ and $p_j^* = \frac{(1-\gamma)\omega}{2} - \frac{(1+\beta)}{2}\bar{p} + \frac{\gamma}{2}\left(p_i^* + \beta\bar{p}\right) + \frac{c}{2} < 0$, such that $\frac{(1-\gamma)\omega}{2} - \frac{(1+\beta)}{2}\bar{p} + \frac{\gamma}{2}\left(p_i^* + \beta\bar{p}\right) + \frac{c}{2} < \frac{(1-\gamma)\omega}{2} - \frac{(1+\beta)}{2}\bar{p} + \frac{\gamma}{2}\beta\bar{p} + \frac{c}{2}$, which yields a contradiction. Therefore, again, only symmetric combinations $p_i^* = p_j^* = p^*$ can constitute mutual best responses and there exists exactly one such combination with $p^* = 0$ if and only if $p_i^*(0) = 0$, i.e., $\frac{(1-\gamma)\omega}{2} - \frac{1+\beta}{2}\bar{p} + \frac{\gamma}{2}\beta\bar{p} + \frac{c}{2} \leq 0 \iff \bar{p} \geq \frac{(1-\gamma)\omega+c}{1+(1-\gamma)\beta} = \bar{p}_L$. Otherwise, the intersection of the best-response functions is determined by $p = \frac{(1-\gamma)\omega}{2} - \frac{1+\beta}{2}\bar{p} + \frac{\gamma}{2}\left(p + \beta\bar{p}\right) + \frac{c}{2} \iff p = \frac{(1-\gamma)\omega+c-\left[1+(1-\gamma)\beta\right]\bar{p}}{(2-\gamma)}$.

For $\bar{p}_L \leq \bar{p} \leq \bar{p}_H$, the equilibrium candidate is $(p,\hat{p}) = (0,\bar{p})$ in case (i) as well as in case (ii) and, hence, unique. If $\bar{p} < \bar{p}_L$, the candidate $(p,\hat{p}) = (0,\bar{p})$ in case (i) was feasible also in case (ii) but not selected. Accordingly, the candidate from case (ii), $(p,\hat{p}) = (\frac{(1-\gamma)\omega+c-\left[1+(1-\gamma)\beta\right]\bar{p}}{(2-\gamma)}, \bar{p})$, is the unique equilibrium candidate. Likewise, for $\bar{p} > \bar{p}_H$, the candidate $(p,\hat{p}) = (0,\bar{p})$ in case (ii) was feasible also in case (i) but not selected. It follows that the candidate from case (i), $(p,\hat{p}) = (0, \frac{(1-\gamma)\omega+\beta c}{(2-\gamma)\beta})$, is the unique equilibrium candidate. Hence, for any $\bar{p}$ and each of the two cases (i) $p^* = 0$ and (ii) $\hat{p}^* = \bar{p}$, we have derived the unique equilibrium candidate.

Step 4: Finally, we need to check that no profitable deviation from the equilibrium candidates exists. From Step 1, we know that the most profitable deviation $(p_i^d, \hat{p}_i^d)$ entails either $p_i^d = 0$ or $\hat{p}_i^d = \bar{p}$.

If $\bar{p} < \bar{p}_L$, the equilibrium candidate $(p, \hat{p}) = (\frac{(1-\gamma)\omega + c - [1+(1-\gamma)\beta]\bar{p}}{(2-\gamma)}, \bar{p})$ was derived by maximizing π_i over p_i given that $\hat{p}_i = \bar{p}$. Accordingly, no profitable deviation with $\hat{p}_i^d = \bar{p}$ can exist. Therefore, a profitable deviation needs to entail $p_i^d = 0$ with additional price being given by (2.A.3). As the equilibrium candidate entails $p_j > 0$ and $\hat{p}_j = \bar{p}$, it holds that $\frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\bar{p} + \frac{\gamma}{2\beta}p_j + \frac{c}{2} > \frac{(1-\gamma)\omega}{2\beta} + \frac{\gamma}{2}\bar{p} + \frac{c}{2} > \bar{p}$ if $\bar{p} < \bar{p}_L$, which implies $\hat{p}_i^d = \bar{p}$. As we have shown that no profitable deviation with $\hat{p}_i^d = \bar{p}$ exists, we can conclude that no profitable deviation exists.

If $\bar{p}_L \leq \bar{p} \leq \bar{p}_H$, the unique equilibrium candidate is given by $(p, \hat{p}) = (0, \bar{p})$. The derivation has shown that no firm has an incentive to deviate (i) to an interior headline price given $\hat{p} = \bar{p}$ and (ii) to an interior additional price given $p = 0$. Accordingly, no profitable deviation exists.

If $\bar{p} > \bar{p}_H$, the equilibrium candidate $(p, \hat{p}) = (0, \frac{(1-\gamma)\omega + \beta c}{(2-\gamma)\beta})$ was derived by maximizing π_i over $\hat{p}_i$ given $p_i = 0$. Accordingly, no profitable deviation with $p_i = 0$ exists. Therefore, a profitable deviation needs to entail $\hat{p}_i^d = \bar{p}$ with the headline price being given by (2.A.6). As the equilibrium candidate entails $p_j = 0$ and $\hat{p}_j < \bar{p}$, it holds that $\frac{(1-\gamma)\omega}{2} - \frac{1+\beta}{2}\bar{p} + \frac{\gamma}{2}\beta\hat{p}_j + \frac{c}{2} < \frac{(1-\gamma)\omega}{2} - \frac{1+\beta}{2}\bar{p} + \frac{\gamma}{2}\beta\bar{p} + \frac{c}{2} < 0$ if $\bar{p} > \bar{p}_H$, which implies $p_i^d = 0$. As we have shown that no profitable deviation with $p_i^d = 0$ exists, we can conclude that no profitable deviation exists. □

Proof of Proposition 2.4: First, we will show that full consumer protection maximizes consumer surplus. First, note that case (i), i.e. $\bar{p} < \bar{p}_L$ in Proposition 2.3 yields the same prices and quantities as Proposition 2.1. Therefore, the comparative statics remain unchanged such that $\frac{\partial \mathscr{C}\mathscr{S}^*}{\partial \beta} > 0$ for $\bar{p} < \bar{p}_L$.

Second, consider case (ii), i.e. $\bar{p} \in [\bar{p}_L, \bar{p}_H]$, now. Plugging the equilibrium prices and quantities into the consumer surplus results in

$$\mathscr{C}\mathscr{S}^* = 2\omega q^* - \left(1+\gamma\right)q^{*2} - 2q^*\bar{p} = \frac{\omega^2 - 2\omega\bar{p} + \beta\left(2-\beta\right)\bar{p}^2}{1+\gamma}.$$

Taking the derivative with respect to β yields $\frac{\partial \mathscr{CS}^*}{\partial \beta} = \frac{2(1-\beta)\bar{p}^2}{1+\gamma} > 0$ such that $\frac{\partial \mathscr{CS}^*}{\partial \beta} > 0$ for $\bar{p} \in [\bar{p}_L, \bar{p}_H]$.

Consider case (iii), i.e. $\bar{p} > \bar{p}_H$, now. Plugging equilibrium quantities into consumer surplus results in

$$\mathscr{CS}^* = 2\omega q^* - (1+\gamma) q^{*2} - 2q^* \hat{p}^* = \frac{\omega^2 - 2\omega\hat{p}^* + \beta(2-\beta)\hat{p}^{*2}}{1+\gamma}.$$

Taking the derivative with respect to β yields

$$\frac{\partial \mathscr{CS}^*}{\partial \beta} = \frac{-2\omega \frac{\partial \hat{p}^*}{\partial \beta} + 2\beta(2-\beta)\hat{p}^* \frac{\partial \hat{p}^*}{\partial \beta} + 2(1-\beta)\hat{p}^{*2}}{1+\gamma}.$$

Inserting $\hat{p}^*$ and $\frac{\partial \hat{p}^*}{\partial \beta}$ into the derivative and rearranging terms yields $\frac{\partial \mathscr{CS}^*}{\partial \beta} > 0 \iff (1-\gamma)(\omega^2 - \beta^2\omega c) + (1-\beta)\beta^2 c^2 > 0$, which is always the case since $c < \omega$. Therefore, we can conclude that $\frac{\partial \mathscr{CS}^*}{\partial \beta} > 0$ for $\bar{p} > \bar{p}_H$. We have shown that $\frac{\partial \mathscr{CS}^*}{\partial \beta} > 0$ in all three cases. As consumer surplus is a continuous function of β, we can conclude that it is monotonically increasing in β. It follows that for all values of $\bar{p}$, consumer surplus is maximized at $\beta = 1$.

To show that consumer surplus is maximized with $\bar{p} = 0$, we will derive the upper bound that maximizes consumer surplus in each of the cases in Proposition 2.3, and then compare these candidates against each other.

Note that for $\bar{p} < \bar{p}_L$, i.e., case (i) of Proposition 2.3, equilibrium prices and quantities are equal to equilibrium prices and quantities in Proposition 2.1. Hence, it remains to hold that $\frac{\partial \mathscr{CS}^*}{\partial \bar{p}} < 0$ for $\bar{p} < \bar{p}_L$. Therefore, the only candidate for the consumer optimal upper bound on the interval $[0, \bar{p}_L]$ is given by $\bar{p} = 0$.

Consider case (ii) now. In equilibrium consumer surplus for $\bar{p} \in [\bar{p}_L, \bar{p}_H]$ equals

$$\mathscr{CS}^* = 2\omega q^* - (1+\gamma) q^{*2} - 2q^* \bar{p} = \frac{\omega^2 - 2\omega\bar{p} + \beta(2-\beta)\bar{p}^2}{1+\gamma}.$$

Taking the derivative with respect to $\bar{p}$ yields $\frac{\partial \mathscr{CS}^*}{\partial \bar{p}} = \frac{-2\omega + 2\beta(2-\beta)\bar{p}}{1+\gamma}$ such that $\frac{\partial \mathscr{CS}^*}{\partial \bar{p}} < 0$ if and only if $\bar{p} < \frac{\omega}{\beta(2-\beta)}$. Depending on parameter constellations, it may hold that $\bar{p}_L < \frac{\omega}{\beta(2-\beta)} < \bar{p}_H$. Consequently, $\bar{p}_L$ and $\bar{p}_H$ are the two potential candidates for the consumer surplus optimal upper bound in the interval $[\bar{p}_L, \bar{p}_H]$.

Lastly, consider case (iii). As neither equilibrium quantities nor prices are affected by a change in $\bar{p}$ in this case, equilibrium consumer surplus is not affected by a change in $\bar{p}$ either. Hence, the optimal upper bound from a consumer surplus perspective is given by $\bar{p}_H$.

As we have identified a unique candidate for each of the three cases, we may now compare the resulting consumer surplus $\mathscr{CS}^*\left(\bar{p}\right)$. As consumer surplus is a continuous function of $\bar{p}$ and strictly decreasing on the interval $[\,0,\bar{p}_L\big)$, $\mathscr{CS}^*(0) > \mathscr{CS}^*\left(\bar{p}_L\right)$ follows. To prove the claim of the proposition, it suffices to show $\mathscr{CS}^*(0) > \mathscr{CS}^*\left(\bar{p}_H\right)$. Note that $c < \bar{p}_L$ and $q^*\left(\bar{p}=c\right) = q^*\left(\bar{p}=\bar{p}_H\right)$. The perceived total prices are identical for these two upper bounds such that consumers demand identical quantities. From the facts that (i) perceived total prices $p^* + \beta\hat{p}^*$ are identical and (ii) the equilibrium additional price is strictly larger with $\bar{p}=\bar{p}_H$ than with $\bar{p}=c$, it follows that actual total prices $p^*+\hat{p}^*$ are higher with $\bar{p}=\bar{p}_H$ than with $\bar{p}=c$. To see this, consider two tuple of prices $(p',\hat{p}')$ and $(p'',\hat{p}'')$ with $p'+\beta\hat{p}' = p''+\beta\hat{p}''$ and $\hat{p}' < \hat{p}''$. Then $p'+\hat{p}' = p''+\beta\hat{p}''+(1-\beta)\hat{p}' < p''+\hat{p}''$. As consumers end up paying less for the same quantities with $\bar{p}=c$ than with $\bar{p}=\bar{p}_H$, it follows that $\mathscr{CS}^*(0) > \mathscr{CS}^*(c) > \mathscr{CS}^*\left(\bar{p}_H\right)$. As we have shown that consumer surplus is maximized with $\beta=1$ as well as with $\bar{p}=0$, we can conclude that full consumer protection maximizes consumer surplus.

Next, we will show that no consumer protection maximizes producer surplus. First, by the same argument as before, the comparative statics result from Proposition 2.2 carry over for $\bar{p} < \bar{p}_L$ such that $\frac{\partial \mathscr{PS}^*}{\partial \beta} < 0$.

Plugging equilibrium prices and quantities of case (ii) into the equation for producer surplus results in

$$\mathscr{PS}^* = 2\frac{\omega-\beta\bar{p}}{1+\gamma}\cdot\left(\bar{p}-c\right).$$

Taking the derivative with respect to β yields $\frac{\partial \mathscr{PS}^*}{\partial \beta} = \frac{-2\bar{p}\left(\bar{p}-c\right)}{1+\gamma} < 0$ such that $\frac{\mathscr{PS}^*}{\partial \beta} < 0$ for $\bar{p} \in [\bar{p}_L,\bar{p}_H]$.

Using the firms' first-order conditions, producer surplus in case (iii) is given by

$$\mathscr{PS}^* = \sum_{i=1}^{2}\pi_i(p^*,\hat{p}^*) = \frac{2(1-\gamma^2)}{\beta}q^{*2},$$

From $\frac{\partial q^*}{\partial \beta} < 0$ it follows that $\frac{\partial \mathscr{P}\mathscr{S}^*}{\partial \beta} = -\frac{2(1-\gamma^2)}{\beta^2} q^{*2} + 4q^* \frac{\partial q^*}{\partial \beta} \frac{1-\gamma^2}{\beta} < 0$. Therefore, we can conclude that $\frac{\partial \mathscr{P}\mathscr{S}^*}{\partial \beta} < 0$ for $\bar{p} > \bar{p}_H$. As $\mathscr{P}\mathscr{S}^*$ is a continuous function of β, it follows that $\mathscr{P}\mathscr{S}^*$ is monotonically decreasing in β and maximized with $\beta = 0$.

Again, the comparative statics result from Proposition 2.2 carry over for $\bar{p} < \bar{p}_L$ such that $\frac{\partial \mathscr{P}\mathscr{S}^*}{\partial \bar{p}} > 0$.

Plugging equilibrium prices and quantities of case (ii) into the equation for producer surplus results in

$$\mathscr{P}\mathscr{S}^* = 2\frac{\omega - \beta\bar{p}}{1+\gamma} \cdot \left(\bar{p} - c\right)$$

Taking the derivative with respect to $\bar{p}$ yields $\frac{\partial \mathscr{P}\mathscr{S}^*}{\partial \bar{p}} = 2\left(\frac{\omega - 2\beta\bar{p} + \beta c}{1+\gamma}\right) > 0 \iff \bar{p} < \frac{\omega + \beta c}{2\beta}$. Note that $\frac{\omega+\beta c}{2\beta} > \bar{p}_H$, such that $\frac{\partial \mathscr{P}\mathscr{S}^*}{\partial \bar{p}} > 0$ for all $\bar{p} \in [\bar{p}_L, \bar{p}_H]$.

Consider the third case now. Since prices and quantities are not affected by a change in the upper bound of additional prices, producer surplus is unaffected by an increase of $\bar{p}$ whenever $\bar{p} > \bar{p}_H$. As $\mathscr{P}\mathscr{S}^*$ is a continuous function of $\bar{p}$, it follows that $\mathscr{P}\mathscr{S}^*$ is monotonically increasing in $\bar{p}$ and maximized with $\bar{p} \to \infty$. As we have shown that producer surplus is maximized with $\beta = 0$ as well as with $\bar{p} \to \infty$, we can conclude that no consumer protection maximizes producer surplus.

Finally, we will derive the welfare results. Equilibrium welfare $\mathscr{W}^*$ is continuous in β as it is continuous in equilibrium quantities, which in turn are continuous in β. Regarding equilibrium quantities $q^*(\beta)$, it holds that $q^*(1) = \frac{\omega - c}{(1+\gamma)(2-\gamma)} < q^{FB}$. Note that $\frac{\partial q^*}{\partial \beta} < 0$ for all $\bar{p}$ at which q^* is differentiable. Hence, q^* is a monotonically decreasing function in β for $\beta \in [0,1]$. Accordingly, there exists a $\beta^{SB} \in (0,1)$ such that $q^*\left(\beta^{SB}\right) = q^{FB}$ if and only if $q^*(0) > q^{FB}$. It holds that $q^*(0) = \min\left\{\frac{\omega - c + \bar{p}}{(1+\gamma)(2-\gamma)}, \frac{\omega}{1+\gamma}\right\} > q^{FB} \iff \bar{p} > \left(1-\gamma\right)(\omega - c)$.

Equilibrium welfare $\mathscr{W}^*$ is continuous in $\bar{p}$ as it is continuous in equilibrium quantities, which in turn are continuous in $\bar{p}$. Regarding equilibrium quantities $q^*(\bar{p})$, it holds that $q^*(0) = \frac{\omega - c}{(1+\gamma)(2-\gamma)} < q^{FB}$. Furthermore $\frac{\partial q^*}{\partial \bar{p}} > 0 \iff \bar{p} < \bar{p}_L$. Hence, there exists at least one $\bar{p}^{SB} > 0$ with $q^*\left(\bar{p}^{SB}\right) = q^{FB}$ if and only if $q^*(\bar{p}_L) \geq q^{FB} \iff \beta \leq \frac{c}{(1-\gamma)\omega + \gamma c}$. □

Proof of Lemma 2.2. First note that by Proposition 2.4, regarding the equilibrium quantities $q^*(\beta)$ it holds that $q^*(0) < q^{FB} \iff \bar{p} \leq (1-\gamma)(\omega - c)$. As q^* is monotonically decreasing in β and equilibrium welfare is concave in quantities, it follows that $\beta^{SB}(\bar{p}) = 0$ if $\bar{p} \leq (1-\gamma)(\omega - c)$. Hence, $\frac{\partial \beta^{SB}(\bar{p})}{\partial \bar{p}} = 0$ for $\bar{p} \leq (1-\gamma)(\omega - c)$.

Now, suppose $\bar{p} > (1-\gamma)(\omega - c)$. In this case, Proposition 2.4 reveals that there exists a unique welfare-maximizing degree of price transparency $\beta^{SB}(\bar{p})$ that induces first-best.

If $\bar{p} < \bar{p}_L$, or equivalently, $\beta < \frac{(1-\gamma)\omega + c - \bar{p}}{(1-\gamma)\bar{p}} \equiv \beta_L$, $\beta^{SB}(\bar{p})$ is implicitly defined by

$$\frac{\omega - c}{1+\gamma} = \frac{\omega - c + (1 - \beta_I^{SB}(\bar{p}))\bar{p}}{(1+\gamma)(2-\gamma)} \quad \Leftrightarrow \quad \beta_I^{SB}(\bar{p}) = 1 - \frac{(1-\gamma)(\omega - c)}{\bar{p}},$$

which is feasible if and only if $\beta_I^{SB} < \beta_L \iff \bar{p} < (1-\gamma)\omega + \gamma c$.

If $\bar{p}_L \leq \bar{p} \leq \bar{p}_H$, or equivalently, $\beta_L \leq \beta \leq \frac{(1-\gamma)\omega}{(2-\gamma)\bar{p} - c} \equiv \beta_H$, $\beta^{SB}(\bar{p})$ is implicitly defined by

$$\frac{\omega - c}{1+\gamma} = \frac{\omega - \beta_{II}^{SB}(\bar{p})\bar{p}}{1+\gamma} \qquad \Leftrightarrow \qquad \beta_{II}^{SB}(\bar{p}) = \frac{c}{\bar{p}}.$$

It needs to hold that $\beta_L \leq \beta_{II}^{SB}(\bar{p}) \leq \beta_H$, which is true if and only if $\bar{p} \geq (1-\gamma)\omega + \gamma c$ and $\bar{p}\left[(2-\gamma)c - (1-\gamma)\omega\right] \leq c^2$. Note that the last inequality holds if either $\frac{\omega}{c} \geq \frac{2-\gamma}{1-\gamma}$ or $\frac{\omega}{c} < \frac{2-\gamma}{1-\gamma}$ and $\bar{p} \leq \frac{c^2}{(2-\gamma)c - (1-\gamma)\omega}$.

If $\bar{p} > \bar{p}_H$, or equivalently $\beta > \beta_H$, $\beta^{SB}(\bar{p})$ is implicitly defined by

$$\frac{\omega - c}{1+\gamma} = \frac{\omega - \beta_{III}^{SB}(\bar{p})c}{(1+\gamma)(2-\gamma)} \qquad \Leftrightarrow \qquad \beta_{III}^{SB}(\bar{p}) = (2-\gamma) - (1-\gamma)\frac{\omega}{c}.$$

It needs to hold that $\beta_{III}^{SB}(\bar{p}) > \beta_H$, which is true if and only if $\frac{\omega}{c} < \frac{2-\gamma}{1-\gamma}$ and $\bar{p} > \frac{c^2}{(2-\gamma)c - (1-\gamma)\omega}$.

We have shown that $\beta^{SB}(\bar{p}) = \beta_I^{SB}(\bar{p})$ such that $\frac{\partial \beta^{SB}(\bar{p})}{\partial \bar{p}} > 0$ if $\bar{p} < (1-\gamma)\omega + \gamma c$. Furthermore, $\beta^{SB}(\bar{p}) = \beta_{II}^{SB}(\bar{p})$ such that $\frac{\partial \beta^{SB}(\bar{p})}{\partial \bar{p}} < 0$ if $\bar{p} \geq (1-\gamma)\omega + \gamma c$ and $\bar{p}\left[(2-\gamma)c - (1-\gamma)\omega\right] \leq c^2$. Finally, $\beta^{SB}(\bar{p}) = \beta_{III}^{SB}(\bar{p})$ such that $\frac{\partial \beta^{SB}(\bar{p})}{\partial \bar{p}} = 0$ if $\frac{\omega}{c} < \frac{2-\gamma}{1-\gamma}$ and $\bar{p} > \frac{c^2}{(2-\gamma)c - (1-\gamma)\omega}$. □

Proof of Lemma 2.3. The proof of Proposition 2.4 has revealed that $q^*(\bar{p}) < q^{FB}$ for all values of $\bar{p}$ if $\beta > \frac{c}{(1-\gamma)\omega + \gamma c} \equiv \beta''$ and that equilibrium quantities are maximized

if $\bar{p} = \bar{p}_L$. As equilibrium welfare is concave in $q^*(\bar{p})$, $\bar{p}^{SB}(\beta) = \bar{p}_L$. Hence, $\frac{\partial \bar{p}^{SB}(\beta)}{\partial \beta} = \frac{\partial \bar{p}_L}{\partial \beta} < 0$ if $\beta \geq \beta''$.

If $\beta < \beta''$, it follows from the proof of Proposition 2.4 that there exists at least one $\bar{p}$ that induces first-best welfare.

The following four observations will help to prove the remaining claims of the lemma:

(i) $\frac{\partial q^*}{\partial \bar{p}} > 0 \Leftrightarrow \bar{p} < \bar{p}_L$, $\frac{\partial q^*}{\partial \bar{p}} < 0 \Leftrightarrow \bar{p}_L < \bar{p} < \bar{p}_H$, and $\frac{\partial q^*}{\partial \bar{p}} = 0 \Leftrightarrow \bar{p} > \bar{p}_H$.

(ii) $q^*(\bar{p} = 0) = \frac{\omega - c}{(1+\gamma)(2-\gamma)} < q^{FB}$.

(iii) q^* strictly decreases in β for $\bar{p} > 0$.

(iv) $q^*(c) = q^*(\bar{p}_H)\ \forall \beta$.

(v) $q^*(c) = q^{FB} \Leftrightarrow \beta = (2-\gamma) - (1-\gamma)\frac{\omega}{c} \equiv \beta' < \beta''$.

First, by observation (i) there exist either one, two, or infinitely many $\bar{p}^{SB}(\beta)$ that induce first-best welfare if $\beta < \beta''$, as equilibrium quantities have to be equal to the first-best quantities to induce first-best welfare.

Second, let $\beta \in [0, \beta')$. Due to observations (ii) and (v) we know that $q^*(0) < q^{FB} < q^*(c)$ in this case. By observation (i) and the fact that $c < \bar{p}_L$ this implies that there can only be one intersection between the first-best quantities and the equilibrium quantities. The resulting unique welfare-maximizing upper bound is implicitly defined by

$$\frac{\omega - c + (1-\beta)\bar{p}^{SB}(\beta)}{(1+\gamma)(2-\gamma)} = \frac{\omega - c}{1+\gamma} \quad \Leftrightarrow \quad \bar{p}^{SB}(\beta) = \frac{(1-\gamma)(\omega - c)}{(1-\beta)}. \tag{2.A.11}$$

Its derivative is given by

$$\frac{\partial \bar{p}^{SB}}{\partial \beta} = \frac{(1-\gamma)(\omega - c)}{(1-\beta)^2} > 0. \tag{2.A.12}$$

Third, let $\beta \in (\beta', \beta'')$. Due to observations (iii), (iv), and (v) it holds that $q^*(c) < q^{FB} < q^*(\bar{p}_L)$ and, equivalently, $q^*(\bar{p}_H) < q^{FB} < q^*(\bar{p}_L)$. Then, observation (i) implies that there are exactly two welfare-maximizing upper bounds, one in case (i) and one in case (ii) of Proposition 2.3. Denote the welfare-maximizing upper

bound with the lower (higher) additional price as $\bar{p}_c^{SB}(\beta)$ ($\bar{p}_p^{SB}(\beta)$). Hence, $\bar{p}_c^{SB}(\beta)$ is implicitly defined by

$$\frac{\omega - c + (1-\beta)\bar{p}_c^{SB}(\beta)}{(1+\gamma)(2-\gamma)} = \frac{\omega - c}{1+\gamma} \quad \Leftrightarrow \quad \bar{p}_c^{SB}(\beta) = \frac{(1-\gamma)(\omega - c)}{(1-\beta)} \tag{2.A.13}$$

such that

$$\frac{\partial \bar{p}_c^{SB}}{\partial \beta} = \frac{(1-\gamma)(\omega - c)}{(1-\beta)^2} > 0. \tag{2.A.14}$$

Similarly, $\bar{p}_p^{SB}(\beta)$ is implicitly defined by

$$\frac{\omega - \beta\bar{p}_p^{SB}(\beta)}{(1+\gamma)} = \frac{\omega - c}{1+\gamma} \quad \Leftrightarrow \quad \bar{p}_p^{SB}(\beta) = \frac{c}{\beta}. \tag{2.A.15}$$

Its derivative is given by

$$\frac{\partial \bar{p}^{SB}}{\partial \beta} = -\frac{c}{\beta^2} < 0. \tag{2.A.16}$$

Fourth, consider the knife-edge case $\beta = \beta'$ now. Due to observation (v), it holds that $q^{FB} = q^*(c)$. We also have that $q^*(c) = q^*(\bar{p})\ \forall \bar{p} \in [\bar{p}_H, \infty)$ due to observation (i) and (iv). This means that the welfare-optimal upper bounds are elements of the set $\{\bar{p} \mid \bar{p} = \bar{p}_c^{SB}(\beta') \vee \bar{p} \geq \bar{p}_H(\beta')\}$.

□

Proof of Proposition 2.5. It holds that $q^*(c) = \frac{\omega - \beta c}{(1+\gamma)(2-\gamma)} = q^*(\bar{p})\ \forall \bar{p} \geq \bar{p}_H$. Accordingly, $\bar{p} = c$ leads to the same level of welfare as all upper bounds $\bar{p} \geq \bar{p}_H$. As perceived total prices are identical but $\bar{p} = c$ leads to an equilibrium with strictly lower additional prices than any $\bar{p} \geq \bar{p}_H$, the resulting total equilibrium prices are strictly lower and, hence, consumer surplus is strictly higher with $\bar{p} = c$ than with $\bar{p} \geq \bar{p}_H$. □

References

Armstrong, Mark, and John Vickers. 2012. "Consumer Protection and Contingent Charges." *Journal of Economic Literature* 50 (2): 477–93. DOI: https://doi.org/10.1257/jel.50.2.477. [73, 83]

Bork, Robert H. 1978. *The Antitrust Paradox : A Policy at War with Itself*. English. Basic Books New York, xi, 462 p. : [69]

Chetty, Raj. 2009. "The Simple Economics of Salience and Taxation." *National Bureau of Economic Research*, DOI: https://doi.org/10.3386/w15246. [70, 72]

Cohen, Mark A. 2012. "Imperfect Competition in Auto Lending: Subjective Markup, Racial Disparity, and Class Action Litigation." *Review of Law & Economics* 8 (1): 21–58. DOI: https://doi.org/10.1515/1555-5879.1501. [68, 69]

de Meza, David, and Diane Reyniers. 2012. "Every Shroud has a Silver Lining: The Visible Benefits of Hidden Surcharges." *Economics Letters* 116 (2): 151–53. DOI: https://doi.org/10.1016/j.econlet.2012.02.027. [72, 74]

Friedman, David Adam. 2020. "Regulating Drip Pricing." *Stanford Law & Policy Review* 31: 51. DOI: https://doi.org/10.2139/ssrn.3337073. [67]

Gabaix, Xavier, and David Laibson. 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *Quarterly Journal of Economics* 121 (2): 505–40. DOI: https://doi.org/10.1162/qjec.2006.121.2.505. [70, 72, 75]

Glaeser, Edward L., and Gergely Ujhelyi. 2010. "Regulating Misinformation." *Journal of Public Economics* 94 (3): 247–57. DOI: https://doi.org/10.1016/j.jpubeco.2010.01.001. [72, 74, 76]

Greenleaf, Eric A., Eric J. Johnson, Vicki G. Morwitz, and Edith Shalev. 2016. "The Price Does Not Include Additional Taxes, Fees, and Surcharges: A Review of Research on Partitioned Pricing." *Journal of Consumer Psychology* 26 (1): 105–24. DOI: https://doi.org/10.1016/j.jcps.2015.04.006. [67]

Grubb, Michael D. 2014. "Consumer Inattention and Bill-Shock Regulation." *Review of Economic Studies* 82 (1): 219–57. DOI: https://doi.org/10.1093/restud/rdu024. [83]

Grunewald, Andreas, Jonathan A Lanning, David C Low, and Tobias Salz. 2020. "Auto Dealer Loan Intermediation: Consumer Behavior and Competitive Effects." Working Paper Series (28136): DOI: https://doi.org/10.3386/w28136. [75]

Heidhues, Paul, Johannes Johnen, and Botond Kőszegi. 2020. "Browsing versus Studying: A Pro-market Case for Regulation." *Review of Economic Studies* 88 (2): 708–29. DOI: https://doi.org/10.1093/restud/rdaa056. [72, 73]

Heidhues, Paul, and Botond Kőszegi. 2018. "Chapter 6 - Behavioral Industrial Organization." In *Handbook of Behavioral Economics - Foundations and Applications 1*. Edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson. Vol. 1, North-Holland, 517–612. DOI: https://doi.org/10.1016/bs.hesbe.2018.07.006. [68, 69]

Heidhues, Paul, Botond Kőszegi, and Takeshi Murooka. 2016a. "Exploitative Innovation." *American Economic Journal: Microeconomics* 8 (1): 1–23. DOI: https://doi.org/10.1257/mic.20140138. [70, 72, 73, 83]

Heidhues, Paul, Botond Kőszegi, and Takeshi Murooka. 2016b. "Inferior Products and Profitable Deception." *Review of Economic Studies* 84 (1): 323–56. DOI: https://doi.org/10.1093/restud/rdw037. [70, 72, 73, 83]

Heyer, Ken. 2006. *Welfare Standards and Merger Analysis: Why Not the Best?* US Department of Justice, Antitrust Division. DOI: https://doi.org/10.2139/ssrn.959454. [69, 94]

Hovenkamp, Herbert. 2012. "Implementing Antitrust's Welfare Goals." *Fordham L. Rev.* 81: 2471. DOI: https://doi.org/10.2139/ssrn.2154499. [69]

Kosfeld, Michael, and Ulrich Schüwer. 2016. "Add-on Pricing in Retail Financial Markets and the Fallacies of Consumer Education." *Review of Finance* 21 (3): 1189–216. DOI: https://doi.org/10.1093/rof/rfw051. [73]

Mohammed, Rafi. 2019. "It's Time to Ban Hidden Fees." *Harvard Business Review*, (2): 2–5. URL: https://hbr.org/2019/02/its-time-to-ban-hidden-fees. [67]

Morwitz, Vicki G., Eric A. Greenleaf, and Eric J. Johnson. 1998. "Divide and Prosper: Consumers' Reactions to Partitioned Prices." *Journal of Marketing Research* 35 (4): 453–63. DOI: https://doi.org/10.1177/002224379803500404. [67]

Office of Fair Trading. 2010. "Advertising of Prices." URL: https://cutt.ly/ahLjncS. OFT1291. [67]

Singh, Nirvikar, and Xavier Vives. 1984. "Price and Quantity Competition in a Differentiated Duopoly." *RAND Journal of Economics* 15 (4): 546–54. URL: http://www.jstor.org/stable/2555525. [69, 74, 93, 94]

Voester, Johannes, Bjoern Ivens, and Alexander Leischnig. 2017. "Partitioned Pricing: Review of the Literature and Directions for Further Research." *Review of Managerial Science* 11 (4): 879–931. DOI: https://doi.org/10.1007/s11846-016-0208-x. [67]

Wilson, Christine S. 2019. "Welfare Standards Underlying Antitrust Enforcement: What You Measure is What You Get." In *Luncheon Keynote Address delivered at the George Mason Law Review 22nd Annual Antitrust Symposium, Arlington, VA.* [69, 94]

Chapter 3

Correlation Neglect, Incentives, and Welfare*

Joint with Matthias Kräkel

3.1 Introduction

In practice, a principal often has more than one performance signal about his agent's effort choice. Typically, the principal observes the realized output of the agent but also receives additional information on the agent's exerted effort. For example, a firm observes the realized sales of its sales agents, but also gets feedback from the customers about the agents' services. An industrial researcher produces output in the form of patents but also receives a performance appraisal by his laboratory head, who observes the researcher's daily work. Many managers do realize not only short-term output but also long-term one, and their exerted effort influences both kinds of output (e.g., quarterly result and annual profit). The common ground of all these examples is that the different performance signals are positively correlated. For example, the sales agent with the highest realized sales is typically also the one with the best customer evaluations.

There exist several studies documenting that real decision makers underestimate the true positive correlation between informative signals, i.e., they suffer from

* Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866. Financial support by the DFG, grant KR 2077/3-1, is gratefully acknowledged. Declarations of interests: none.

correlation neglect, e.g., Ortoleva and Snowberg (2015), Eyster and Weizsacker (2016), Enke and Zimmermann (2019), and Rees-Jones, Shorrer, and Tergiman (2020). Concerning our principal-agent framework, this means that an agent that suffers from correlation neglect underestimates the true correlation between the performance measures that are used by the principal to generate incentives.[1]

In this note, we modify the well-known model by Holmström and Milgrom (1987, 1990, 1991) to analyze how correlation neglect influences the principal's optimal incentive contract and material welfare. For this purpose, we assume that the principal can make use of two performance measures about the agent's effort choice and that the agent underestimates the positive correlation between the two measures. At first glance, one might conjecture that correlation neglect leads to higher implemented effort and, therefore, to a less inefficient outcome, because correlation neglect implies a lower perceived risk premium of the agent, which makes effort implementation less costly for the principal. However, our analysis shows that this intuition will only hold if the agent's perceived correlation coefficient is sufficiently small. Otherwise, the principal prefers to put a positive weight on the more precise performance measure and a negative weight on the other one. Such contract reduces incentives and is bad from a welfare perspective but optimal for the principal to generate a first-order advantage in form of better insurance.

3.2 The Model

We modify the principal-agent framework with linear contracts of Holmström and Milgrom (1987, 1990, 1991). The principal (P) has to hire an agent (A) to run a business. If A is hired, he will produce output according to the production function $x = e + \varepsilon$ with $e \geq 0$ denoting effort chosen by A, and ε denoting a normally distributed noise term with mean zero and variance σ_ε^2. Whereas x is observable by P and verifiable by a third party, neither e nor ε are observable by P, who thus faces a moral-hazard problem. Exerting effort leads to personal costs for A that can be

1. Our analysis is based on an underestimation of the true covariance, leading to an underestimation of the true correlation coefficient.

described in monetary terms by the function $c(e) = ke^2/2$ with $k > 0$. We assume that P can make use of a second performance measure, $s = e + \theta$, with θ denoting a normally distributed noise term with mean zero and variance σ_θ^2. Analogously to x, the measure s is observable by P and verifiable by a third party, but e and θ are unobservable. The lines between the two measures are clear and exogenously given. The noise terms are assumed to be positively correlated with $\rho \in (0,1)$ as correlation coefficient. We depart from the Holmström-Milgrom setting by assuming that A suffers from correlation neglect. In our setting, this means that A believes the correlation coefficient to be $\hat{\rho}$ with $\hat{\rho} < \rho$, whereas P knows that the true correlation coefficient is ρ. As is standard in the literature on contracting with behavioral agents (e.g., Kőszegi (2014)), we assume that the parties "agree to disagree", i.e.. principal and agent do not update their beliefs over the course of the game.

Following Holmström and Milgrom, we restrict the class of incentive schemes for the agent to linear ones (see Holmström and Milgrom (1987) for a justification): $w(x,s) = \alpha + \beta \cdot x + \gamma \cdot s$ with α as fixed salary, and β and γ as piece rates. P is assumed to be risk neutral, maximizing expected profits $E[x - w(x,s)]$. However, A is assumed to be risk averse with utility function $1 - \exp\{-r \cdot (w(x,s) - c(e))\}$. Here $r > 0$ denotes the (constant) Arrow-Pratt measure of absolute risk aversion. A maximizes his perceived expected utility. As is well-known for the Holmström-Milgrom model, due to the normally distributed noise and the linearity assumptions, this is equivalent to maximizing the certainty equivalent

$$CE = \alpha + \beta \cdot e + \gamma \cdot e - \frac{k}{2}e^2 - \frac{r}{2}\left(\beta^2\sigma_\varepsilon^2 + \gamma^2\sigma_\theta^2 + 2\hat{\rho}\sigma_\varepsilon\sigma_\theta\gamma\beta\right). \tag{3.1}$$

Due to correlation neglect, the risk premium of the certainty equivalent is based on A's perceived correlation coefficient, $\hat{\rho}$. W.l.o.g., A's reservation value is normalized to zero. The timing is the usual one: 1. P offers $w(x,s)$ to A; 2. A accepts or rejects the offer; 3. A chooses e in case of acceptance; 4. x and s are realized and payments are made. As a reference solution, we can consider the efficient or first-best outcome, which will be realized if P perfectly observes effort e and can write forcing contracts without any transaction costs. In that case, P would implement $e^{FB} = 1/k$ to maximize the first-best material welfare $E[x] - c(e)$.

3.3 Solution to the Model

The game is solved by backward induction, starting with stage 3 in which A chooses effort to maximize CE. As, according to (3.1), CE is strictly concave, optimal effort is described by the first-order condition, leading to the incentive constraint $e^* := (\beta + \gamma)/k$. In stage 1, P offers an optimal contract that extracts all rents from A.[2] In other words, the participation constraint binds under the optimal contract so that P chooses $w(x,s)$ to maximize the *perceived* second-best welfare that is based on $\hat{\rho}$ because P only has to compensate A for the *perceived* risk premium. Hence, P solves

$$\max_{\beta,\gamma} \pi \;=\; \max_{\beta,\gamma} \left(e^* - \frac{k}{2} e^{*2} - \frac{r}{2} \left(\beta^2 \sigma_\varepsilon^2 + \gamma^2 \sigma_\theta^2 + 2\hat{\rho}\sigma_\varepsilon \sigma_\theta \gamma \beta \right) \right) =: \; \pi^*. \quad (3.2)$$

As the respective Hessian matrix is negative semi-definite,[3] the optimal piece rates, $\beta^*(\hat{\rho})$ and $\gamma^*(\hat{\rho})$, are described by the first-order conditions:

$$\beta^*(\hat{\rho}) = \left(\frac{\sigma_\theta}{\sigma_\varepsilon} - \hat{\rho} \right) \frac{\sigma_\theta \sigma_\varepsilon}{\Psi} \quad \text{and} \quad \gamma^*(\hat{\rho}) = \left(\frac{\sigma_\varepsilon}{\sigma_\theta} - \hat{\rho} \right) \frac{\sigma_\theta \sigma_\varepsilon}{\Psi} \quad (3.3)$$

with $\Psi := \sigma_\theta^2 + \sigma_\varepsilon^2 - 2\hat{\rho}\sigma_\theta\sigma_\varepsilon + kr\sigma_\theta^2\sigma_\varepsilon^2(1-\hat{\rho}^2) > 0$. The two expressions in (3.3) show that either both optimal piece rates are positive or one piece rate is positive, and the other one negative. For example, whenever the two performance measures are equally precise, i.e., if $1/\sigma_\theta^2 = 1/\sigma_\varepsilon^2 \Leftrightarrow \sigma_\theta = \sigma_\varepsilon$, both optimal piece rates are positive. However, in all other cases, i.e., $\sigma_\theta \neq \sigma_\varepsilon$, the sign of the optimal piece rates depends on the perceived correlation $\hat{\rho}$. In particular, if $\hat{\rho}$ is sufficiently low, then $\beta^*(\hat{\rho}), \gamma^*(\hat{\rho}) > 0$. If, however, $\hat{\rho}$ is sufficiently high, the optimal piece rate that corresponds to the more precise performance measure will be positive and the other one negative (e.g., $\beta^*(\hat{\rho}) < 0$ and $\gamma^*(\hat{\rho}) > 0$ can only be possible if $\sigma_\theta < \sigma_\varepsilon$, i.e., if the precision $1/\sigma_\theta^2$ is higher than the precision $1/\sigma_\varepsilon^2$). The two possible cases are quite intuitive. Consider the perceived optimal risk premium, $\frac{r}{2}(\beta^*(\hat{\rho})^2 \sigma_\varepsilon^2 + \gamma^*(\hat{\rho})^2 \sigma_\theta^2 + 2\hat{\rho}\sigma_\varepsilon\sigma_\theta\gamma^*(\hat{\rho})\beta^*(\hat{\rho}))$, which consists of two idiosyncratic parts, $\beta^*(\hat{\rho})^2 \sigma_\varepsilon^2$ and $\gamma^*(\hat{\rho})^2 \sigma_\theta^2$, and a joint part, $2\hat{\rho}\sigma_\varepsilon\sigma_\theta\gamma^*(\hat{\rho})\beta^*(\hat{\rho})$. The

2. This is due to the assumption of unlimited liability.

3. I.e., $\partial^2\pi/\partial\beta^2 = -(kr\sigma_\varepsilon^2 + 1)/k < 0$, $\partial^2\pi/\partial\gamma^2 = -(kr\sigma_\theta^2 + 1)/k < 0$ as well as $\partial^2\pi/\partial\beta^2 \cdot \partial^2\pi/\partial\gamma^2 - \partial^2\pi/\partial\beta\partial\gamma = r[kr\sigma_\theta^2\sigma_\varepsilon^2(1-\rho^2) + \sigma_\theta^2 + \sigma_\varepsilon^2 - 2\rho\sigma_\theta\sigma_\varepsilon]/k > 0$.

principal P wants to implement a specific effort at minimum risk costs. If $\hat{\rho}$ is sufficiently small, the joint part of the risk premium will be less important than the idiosyncratic parts and P chooses two positive values for the optimal piece rates to create incentives (note that $\beta^*\left(\hat{\rho}\right)$ and $\gamma^*\left(\hat{\rho}\right)$ are quadratic in the idiosyncratic parts so that negative piece rates do not make sense). If, however, the correlation $\hat{\rho}$ is sufficiently high, then the reduction of the risk premium should be mainly done via the joint part. Hence, P optimally puts a negative weight on one of the two performance measures, so that $2\hat{\rho}\sigma_\varepsilon\sigma_\theta\gamma\beta < 0$, to reduce the overall risk premium.[4] This effect is similar to the reduction of risk costs via relative performance evaluation of two agents whose tasks are subject to the same stochastic influences (e.g., two sales agents who sell the same kind of product).

Next, we analyze the welfare implications of correlation neglect. The implemented effort under the optimal contract for agents with correlation neglect, $e^*\left(\hat{\rho}\right)$, can be computed by inserting the expressions for $\beta^*\left(\hat{\rho}\right)$ and $\gamma^*\left(\hat{\rho}\right)$ into the incentive constraint $e^* = (\beta + \gamma)/k$. Let $e^*\left(\rho\right)$ denote the respective effort for $\hat{\rho} = \rho$, i.e., the optimal effort that will be implemented by P if A does not suffer from correlation neglect. Both $e^*\left(\hat{\rho}\right)$ and $e^*\left(\rho\right)$ are smaller than e^{FB} because $e^* = (\beta + \gamma)/k$ and $\beta^*(\tau) + \gamma^*(\tau) < 1$ for $\tau = \hat{\rho}, \rho$ (see (3.3)). Thus, due to the moral-hazard problem, P does not implement the effort level that maximizes material welfare, irrespective of whether A suffers from correlation neglect or not. The following proposition describes under which condition correlation neglect is welfare enhancing:

Proposition 3.1. *The optimal effort implemented under correlation neglect, $e^*\left(\hat{\rho}\right)$, will be larger than the optimal effort without correlation neglect, $e^*\left(\rho\right)$, if and only if*

$$\frac{2\sigma_\theta\sigma_\varepsilon}{\sigma_\theta^2 + \sigma_\varepsilon^2} > \frac{\rho^2 - \hat{\rho}^2}{\rho\left(1 - \hat{\rho}^2\right) - \hat{\rho}\left(1 - \rho^2\right)}. \tag{3.4}$$

Condition (3.4) shows that there exist feasible parameter constellations so that both relations $e^*\left(\hat{\rho}\right) > e^*\left(\rho\right)$ and $e^*\left(\hat{\rho}\right) < e^*\left(\rho\right)$ are possible.[5] To get an intuition,

4. As the absolute value of the positive piece rate must be larger than the absolute value of the negative piece rate to induce $e^* > 0$, the positive piece rate is assigned to the more precise performance measure to minimize the two idiosyncratic parts of the perceived risk premium.

5. To see this, consider the parameter constellations $\sigma_\theta = 1$, $\sigma_\varepsilon = 1$, $\rho = 0.5$, $\hat{\rho} = 0$, and $\sigma_\theta = 5$, $\sigma_\varepsilon = 1$, $\rho = 0.5$, $\hat{\rho} = 0$.

we start with the case of a high correlation between the two performance measures, i.e., ρ is close to one. In this case, the optimal contract for rational agents will always implement a higher effort than the optimal contract for an agent who suffers from correlation neglect. The reason is that *P* can provide high incentives and filter almost all the risk perceived by a rational agent with an appropriate combination of a positive and a negative piece rate, whereas this is not possible for an agent who suffers from correlation neglect. Consider the rational agent first. He correctly anticipates the high correlation between the two performance measures. Consequently, he treats a slightly negative piece rate as very effective in reducing the joint part of the risk premium. This allows *P* to implement high effort because he can effectively reduce the perceived risk induced by a large positive piece rate via a negative piece rate for the other performance measure that is relatively small in absolute terms. In the extreme case, when $\rho \to 1$, the principal can even implement an effort level that is arbitrarily close to the first-best effort level.

However, an agent who suffers from correlation neglect considers a negative piece rate as less effective in reducing the perceived risk that is induced by a high positive piece rate. In the extreme case, when $\rho \to 1$, such that *P* filters out almost all risk and provide high incentives for the rational agent, the agent who suffers from correlation neglect still believes that the contract is risky. Thus, *P* implements lower effort by an agent who neglects correlation compared to a rational agent. While a high correlation of the two performance measures leads to a clear-cut comparison between the implemented efforts, the economic intuition for intermediate and low correlation is less clear and can best be interpreted by considering marginal changes in correlation neglect, i.e., a comparative-static analysis of $e^*\left(\hat{\rho}\right)$ with respect to $\hat{\rho}$. We obtain

$$\begin{aligned}\frac{\partial e^*}{\partial \hat{\rho}} &= -2r\sigma_\theta^2\sigma_\varepsilon^2 \frac{\left(\sigma_\varepsilon - \hat{\rho}\sigma_\theta\right)\left(\sigma_\theta - \hat{\rho}\sigma_\varepsilon\right)}{\left(kr\sigma_\theta^2\sigma_\varepsilon^2\left(1-\hat{\rho}^2\right) + \sigma_\theta^2 + \sigma_\varepsilon^2 - 2\hat{\rho}\sigma_\theta\sigma_\varepsilon\right)^2} \\ &= -2r\sigma_\theta^2\sigma_\varepsilon^2 \cdot \beta^*\left(\hat{\rho}\right) \cdot \gamma^*\left(\hat{\rho}\right).\end{aligned}$$

Thus, $sign(\partial e^*/\partial\hat{\rho}) = -sign(\beta^*\left(\hat{\rho}\right) \cdot \gamma^*\left(\hat{\rho}\right))$. Note that $\partial e^*/\partial\hat{\rho}$ denotes the effect of a marginal change in correlation neglect on optimal effort, as it describes the change of the optimal effort if the correlation coefficient marginally decreases.

First, consider the case of $\beta^*(\rho) \cdot \gamma^*(\rho) > 0$ – both optimal piece rates are positive for rational agents – implying $\frac{\partial e^*}{\partial \hat{\rho}}\Big|_{\hat{\rho}=\rho} < 0$. From the considerations above, we know that we are in a situation in which the perceived correlation coefficient, $\hat{\rho}$, is sufficiently small, and the joint part of the perceived risk premium is positive. If in such a situation, $\hat{\rho}$ further decreases, the joint part of the perceived risk premium goes down so that implementing higher effort becomes less costly for *P*. Consequently, *P* will implement a higher effort level if *A* suffers from correlation neglect. *Second*, consider the case of $\beta^*(\rho) \cdot \gamma^*(\rho) < 0$ – one optimal piece rate is positive and the other one negative for rational agents – implying $\frac{\partial e^*}{\partial \hat{\rho}}\Big|_{\hat{\rho}=\rho} > 0$. Now, $\hat{\rho}$ is sufficiently high so that *P* prefers a negative weight for the joint part of the perceived risk premium to reduce overall risk costs. If in such a situation, the perceived correlation becomes lower, the reduction of the perceived risk premium via the joint part will become less effective. As a consequence, *P* prefers to adjust incentives downwards and implements a lower effort level.

The scenario described in Proposition 1 is a little bit more subtle. Either we have one of the two previous cases so that ρ and $\hat{\rho}$ are both low or both intermediate so that the arguments from the comparative-static analysis on marginal changes in correlation go through. Or the degree of correlation neglect, $\rho - \hat{\rho}$, is so large that we switch from the second case (intermediate correlation) to the first case (low correlation). However, as the right-hand side of (3.4) is monotonically increasing in $\hat{\rho}$, the stronger *A* suffers from correlation neglect, the more condition (3.4) tends to hold, so that correlation neglect leads to the implementation of higher effort, whenever ρ is not too large. Intuitively, a switch from the second case to the first case means that *P* chooses two positive piece rates under correlation neglect, which both contribute to higher incentives, whereas *P* prefers one positive and one negative piece rate without correlation neglect.

The left-hand side of (3.4) illustrates the influence of the precision of both performance measures on the implications of correlation neglect. If the precision of both measures is very similar (i.e., $\sigma_\theta \approx \sigma_\varepsilon$), the left-hand side of (3.4) attains its maximum 1 so that the condition is satisfied and correlation neglect leads to

higher effort implementation.[6] This outcome corresponds to optimal piece rates that are both positive due to $\sigma_\theta \approx \sigma_\varepsilon$ (see (3.3)). If both performance measures have very different precisions (i.e., σ_θ and σ_ε differ considerably), the left-hand side of (3.4) will become very small so that the condition tends to be violated. In that case, correlation neglect yields lower optimal effort. Similar to the observations above, sufficiently different values of σ_θ and σ_ε imply that one of the optimal piece rates in (3.3) is negative and the other one positive.

Condition (3.4) has shown that the interplay of the precisions of the two performance measures, the magnitude of the true correlation, and the magnitude of correlation neglect determines whether material welfare becomes larger or smaller under correlation neglect. The following result states sufficient conditions for either outcome:[7]

Proposition 3.2. *Define* $\bar{\rho} := \min\{\frac{\sigma_\theta}{\sigma_\varepsilon}, \frac{\sigma_\varepsilon}{\sigma_\theta}\}$. *(i)* $e^*\left(\rho\right) < e^*\left(\hat{\rho}\right) < e^{FB}$ *if* $\rho \in (0, \bar{\rho}\,]$, *and (ii)* $e^*\left(\hat{\rho}\right) < e^*\left(\rho\right) < e^{FB}$ *if* $\hat{\rho} \in [\bar{\rho}, 1)$.

The proposition shows that if the correlation is sufficiently low and, therefore, the perceived correlation is low as well, correlation neglect will be efficiency-enhancing. However, correlation neglect will lead to lower material welfare if the perceived correlation, and, hence, also the true correlation, is large. The economic intuition for these findings is the following. The discussion above has shown that we can differentiate between two cases for the perceived correlation coefficient. In case of low correlation, both performance measures are positive and, thus, used for incentivizing A, leading to high effort. In case of high correlation, only the more precise performance measure is used to incentivize A, whereas the other one purely serves insurance purposes at the cost of reducing incentives.

We conclude our analysis by considering the implications of correlation neglect on P's and A's expected utilities. We start with P's expected utility, which is given by his expected profit. Let $\pi^*\left(e^*\left(\hat{\rho}\right)\right)$ denote the optimal expected profit under cor-

6. Whereas the left-hand side of (3.4) is equal or smaller than one, the right-hand side is strictly smaller than one.

7. All proofs are relegated to the appendix.

relation neglect, which can be computed by plugging the optimal values for $e^*(\hat{\rho})$, $\beta^*(\hat{\rho})$, and $\gamma^*(\hat{\rho})$ in (3.2). The optimal expected profit without correlation neglect, $\pi^*(e^*(\rho))$, can be computed analogously. We obtain $\pi^*(e^*(\hat{\rho})) = e^*(\hat{\rho})/2$ and $\pi^*(e^*(\rho)) = e^*(\rho)/2$, which leads to the following result:

Proposition 3.3. *The optimal expected profit under correlation neglect, $\pi^*(e^*(\hat{\rho}))$, will be larger than the optimal expected profit without correlation neglect, $\pi^*(e^*(\rho))$, if and only if condition (3.4) is satisfied.*

Again, a comparative-static analysis is helpful in illustrating the economic intuition for our findings. Using π^* described by (3.2), the envelope theorem immediately leads to $d\pi^*/d\hat{\rho} = \partial\pi^*/\partial\hat{\rho} = -r\sigma_\varepsilon\sigma_\theta\gamma^*(\hat{\rho})\beta^*(\hat{\rho})$ so that $sign(d\pi^*/d\hat{\rho}) = -sign(\beta^*(\hat{\rho})\cdot\gamma^*(\hat{\rho}))$. In analogy to above, $d\pi^*/d\hat{\rho}$ can be interpreted as the marginal correlation neglect effect on optimal profit. Thus, we can again differentiate between the two cases considered above. First, suppose that $\beta^*(\hat{\rho})\cdot\gamma^*(\hat{\rho}) > 0$ with $\hat{\rho} = \rho$, implying a positive joint part of the perceived risk premium. If in such a situation *A*'s perceived correlation goes down, the overall perceived risk premium will go down as well so that the optimal expected profit increases due to the binding participation constraint. However, in case of $\beta^*(\hat{\rho})\cdot\gamma^*(\hat{\rho}) < 0$ with $\hat{\rho} = \rho$ the joint part of the perceived risk premium is negative. Now, a perceived decrease of correlation increases the perceived risk premium, for which *A* has to be compensated by *P*, so that *P*'s optimal expected profit is reduced.

A's expected utility is expressed by his certainty equivalent (3.1). In our analysis, *P* extracts all perceived rents from *A*. However, in line with De la Rosa (2011), in the following we consider two ways how *A* could profit or suffer from correlation neglect. *First*, we may assume some bargaining power on the side of the agent. So far, we assumed that *P* has all the bargaining power. Reversing this assumption by introducing Bertrand competition between several identical principals in the labor market would redistribute the complete surplus to *A*. Consequently, *A* profits from correlation neglect if condition (3.4) holds and suffers from his mistaken belief if (3.4) is violated. Obviously, this argument does hold whenever the market makes *P*

and *A* share the surplus of the contract. Notably, under a given sharing rule, *P* and *A* will profit from correlation neglect at the same time if material welfare increases.

Second, de la Rosa (2011) uses the actual expected utility of the agent, i.e. his expected utility under the *true* distribution of outcomes, as a measure of the agent's well-being. This consideration leads to the following proposition:

Proposition 3.4. *The actual expected utility of an agent with correlation neglect will be higher than the expected utility of a rational agent if and only if* $\beta^*\left(\hat{\rho}\right) \cdot \gamma^*\left(\hat{\rho}\right) < 0$.

The proposition shows that agents with correlation neglect receive a positive rent whenever their contract exhibits a positive and a negative piece rate, and their expected utility is evaluated under the true distribution of the performance measures. The reason for this result is that correlation neglect works as a kind of commitment for the agent. More precisely, in the case of $\beta^*\left(\hat{\rho}\right) \cdot \gamma^*\left(\hat{\rho}\right) < 0$, which implies that $\beta^*\left(\rho\right) \cdot \gamma^*\left(\rho\right) < 0$, agents with correlation neglect are committed to reject a contract when it only offers low insurance for the joint risk in the two performance measures. As the principal anticipates this behavior, the agent's commitment forces him to offer a contract with higher insurance for this risk. Under the true distribution, the insurance is even more valuable to the agents, because the joint risk is even higher. Consequently, rational agents would prefer the optimal contract offered to agents with correlation neglect to the contract offered to rational agents.[8] Whenever the two piece rates are positive, the opposite argument holds. The insurance provided for the joint risk in the two performance measures is lower in an optimal contract for agents with correlation neglect than for rational agents such that correlation neglect makes agents worse off under the true distribution.

8. This argumentation is similar to that of other papers, e.g., de la Rosa (2011) and Babaioff, Gonczarowski, and Romm (2019), in which an agent's low sophistication works as a commitment and, therefore, improves his outcomes.

Appendix 3.A Proofs

Proof of Proposition 3.2. (i) From the discussion of Proposition 1, we know that $\partial e^*\left(\tilde{\rho}\right)/\partial\tilde{\rho} < 0$ for all $\tilde{\rho} \in [\hat{\rho}, \rho)$ whenever $\beta^*\left(\rho\right) \cdot \gamma^*\left(\rho\right) \geq 0$. The inequality $\beta^*\left(\rho\right) \cdot \gamma^*\left(\rho\right) \geq 0$ will hold iff $\rho \leq \bar{\rho} = \min\{\frac{\sigma_\theta}{\sigma_\varepsilon}, \frac{\sigma_\varepsilon}{\sigma_\theta}\}$. (ii) Analogously, we also know that $\partial e^*\left(\tilde{\rho}\right)/\partial\tilde{\rho} > 0$ for all $\tilde{\rho} \in (\hat{\rho}, \rho]$ whenever $\beta^*\left(\hat{\rho}\right) \cdot \gamma^*\left(\hat{\rho}\right) \leq 0$. The inequality $\beta^*\left(\hat{\rho}\right) \cdot \gamma^*\left(\hat{\rho}\right) \leq 0$ will hold iff $\hat{\rho} \geq \bar{\rho} = \min\{\frac{\sigma_\theta}{\sigma_\varepsilon}, \frac{\sigma_\varepsilon}{\sigma_\theta}\}$. □

Proof of Proposition 3.4. Note that the certainty equivalent of a rational agent under the optimal contract is equal to zero. Calculating the certainty equivalent of an agent that neglects correlation under the true distribution of performance measures yields

$$\begin{aligned}&\alpha^*\left(\hat{\rho}\right) + \left[\beta^*\left(\hat{\rho}\right) + \gamma^*\left(\hat{\rho}\right)\right] \cdot e^*\left(\hat{\rho}\right) - \frac{k}{2}e^*\left(\hat{\rho}\right)^2 \\ &-\frac{r}{2}\left(\beta^*\left(\hat{\rho}\right)^2\sigma_\varepsilon^2 + \gamma^*\left(\hat{\rho}\right)^2\sigma_\theta^2 + 2\rho\sigma_\varepsilon\sigma_\theta\gamma^*\left(\hat{\rho}\right)\beta^*\left(\hat{\rho}\right)\right),\end{aligned}$$

with $\alpha^*\left(\hat{\rho}\right)$ denoting the agent's fixed salary under the optimal contract. As

$$\begin{aligned}&\alpha^*\left(\hat{\rho}\right) + \left[\beta^*\left(\hat{\rho}\right) + \gamma^*\left(\hat{\rho}\right)\right] \cdot e^*\left(\hat{\rho}\right) - \frac{k}{2}e^*\left(\hat{\rho}\right)^2 \\ &-\frac{r}{2}\left(\beta^*\left(\hat{\rho}\right)^2\sigma_\varepsilon^2 + \gamma^*\left(\hat{\rho}\right)^2\sigma_\theta^2 + 2\hat{\rho}\sigma_\varepsilon\sigma_\theta\gamma^*\left(\hat{\rho}\right)\beta^*\left(\hat{\rho}\right)\right) = 0,\end{aligned}$$

the certainty equivalent of a rational agent under the optimal contract will be smaller than the certainty equivalent of an agent with correlation neglect under the true distribution of the performance measures iff $-\left(\rho - \hat{\rho}\right)\gamma^*\left(\hat{\rho}\right)\beta^*\left(\hat{\rho}\right) > 0$. □

References

Babaioff, Moshe, Yannai A. Gonczarowski, and Assaf Romm. 2019. "Playing on a Level Field: Sincere and Sophisticated Players in the Boston Mechanism with a Coarse Priority Structure." *Proceedings of the 2019 ACM Conference on Economics and Computation*, (06): DOI: https://doi.org/10.1145/3328526.3329596. [118]

de la Rosa, Leonidas Enrique. 2011. "Overconfidence and Moral Hazard." *Games and Economic Behavior* 73 (2): 429–51. DOI: https://doi.org/10.1016/j.geb.2011.04.001. [118]

Enke, Benjamin, and Florian Zimmermann. 2019. "Correlation Neglect in Belief Formation." *Review of Economic Studies* 86 (1): 313–32. DOI: https://doi.org/10.1093/restud/rdx081. [110]

Eyster, Erik, and Georg Weizsacker. 2016. "Correlation Neglect in Portfolio Choice: Lab Evidence." *Discussion Paper*, (10): DOI: https://doi.org/10.2139/ssrn.2914526. [110]

Holmström, Bengt, and Paul Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives." *Econometrica* 55 (2): 303–28. DOI: https://doi.org/10.2307/1913238. [110, 111]

Holmström, Bengt, and Paul Milgrom. 1990. "Regulating Trade Among Agents." *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft* 146 (1): 85–105. DOI: https://doi.org/10.2307/40751306. [110]

Holmström, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, Organization* 7: 24–52. DOI: https://doi.org/10.2307/764957. [110]

Kőszegi, Botond. 2014. "Behavioral Contract Theory." *Journal of Economic Literature* 52 (4): 1075–118. DOI: https://doi.org/10.1257/jel.52.4.1075. [111]

Ortoleva, Pietro, and Erik Snowberg. 2015. "Overconfidence in Political Behavior." *American Economic Review* 105 (2): 504–35. DOI: https://doi.org/10.1257/aer.20130921. [110]

Rees-Jones, Alex, Ran Shorrer, and Chloe J. Tergiman. 2020. "Correlation Neglect in Student-to-School Matching." Working Paper 26734. National Bureau of Economic Research. DOI: https://doi.org/10.3386/w26734. [110]

Chapter 4

Games Between Players with Dual-Selves*

Joint with Simon Dato and Andreas Grunewald

4.1 Introduction

There is a growing consensus that dual-self processes are essential to human decision making. For example, individuals strategically manipulate their beliefs to improve their own subsequent behavior (Bénabou and Tirole, 2006a; Schwardmann, Tripodi, and Weele, 2019; Schwardmann and Weele, 2019), choose actions in order to boost their self-image (Brunnermeier and Parker, 2005; Bénabou and Tirole, 2011), and restrict their (beliefs about) possible future action sets to improve behavior in the short term (O'Donoghue and Rabin, 1999; Fudenberg and Levine, 2006; Brocas and Carrillo, 2008). Nevertheless, we still lack a thorough understanding of how such dual-self processes affect economic decision-making in individual decision environments and even more so in strategic interactions with other players with dual-selves.

This paper takes a step towards facilitating the analysis of decisions by individuals with dual-selves. For this purpose, we conceptualize decision processes in which beliefs and actions follow from different or even opposing objectives and do not necessarily constitute coherent consequences of each other. With dual-selves,

* Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866. Financial support by the DFG, grant KR 2077/3-1, is gratefully acknowledged. Declarations of interests: none.

every decision process is inherently strategic because different selves may follow opposing objectives. Therefore, we study a class of games that allows to capture individual decision environments and enable the analysis of strategic situations. In our setup, each player consists of two selves that interact non-cooperatively, choosing a player's action and beliefs.

We first define the solution concept *Dual-Selves equilibrium*, which can be viewed as a generalization of Nash equilibrium to study Dual-Selves games. The key idea underlying this concept is that, in equilibrium, beliefs are optimally chosen according to the corresponding self's objectives and given all other selves' strategies in the game. Actions then constitute a best response to these beliefs. While our concept thus explicitly allows for multiple players, the definition of a *Dual-Selves equilibrium* is, in fact, also a necessary prerequisite for analyzing individual decisions if there is intra-personal conflict. Our main theorem shows that a *Dual-Selves equilibrium* exists under relatively mild conditions on the two selves' strategic interaction. Hence, no matter whether we consider decisions of only one individual or strategic interactions of several players, *Dual-Selves equilibrium* provides a coherent concept of how to think about decision making.

The conceptualization of decisions by individuals with dual-selves, the definition of a *Dual-Selves equilibrium*, and our existence result pave the way for the analysis of various applications where dual-self processes affect decision-making. Such applications can, for example, comprise effort decisions in promotion tournaments if self-esteem is important, the selection of political platforms if politicians manipulate their beliefs about the state of the world, bargaining situations in which individuals strategically manipulate their beliefs about outside options, and shopping decisions if consumers have self-control problems when facing the goods. The growing empirical evidence on the prevalence of such dual-self processes (see, for example Brocas and Carrillo, 2008; Alonso, Brocas, and Carrillo, 2013; Schwardmann, Tripodi, and Weele, 2019; Schwardmann and Weele, 2019) necessitates a concept that allows researchers to make additional theoretical predictions in individual decision-making and strategic interactions. For this purpose, the *Dual-Selves equilibrium* proposed in this paper serves as a useful concept.

The class of games that we study and the proposed equilibrium concept encompass a variety of applications. While our primary focus is on actual dual-self processes, our framework allows us also to analyze the strategic interaction of single-self players with limited cognitive capabilities or non-standard preferences. In this realm, several solution concepts have been proposed to relax the assumptions forming the basis of Nash equilibrium and its refinements. For instance, they allow players to have misspecified beliefs (Eyster and Rabin, 2005; Jehiel, 2005; Esponda and Pouzo, 2016) or players' utilities to directly depend on beliefs (Geanakoplos, Pearce, and Stacchetti, 1989; Dato, Grunewald, Müller, and Strack, 2017).[1] These solution concepts have successfully explained evidence arising from strategic situations that is hard to reconcile with the notion of Nash equilibrium. We show that games between players with limited cognitive capabilities can be interpreted as Dual-Selves games and that the proposed equilibrium concepts are subcases of *Dual-Selves equilibrium*.

We further contribute to this literature in two ways. First, by analyzing potentially non-finite games, we extend the previously established existence results that are restricted to finite games (e.g., Geanakoplos, Pearce, and Stacchetti, 1989; Eyster and Rabin, 2005; Esponda and Pouzo, 2016; Dato et al., 2017). Showing existence in non-finite games allows us to analyze how limited cognitive capabilities affect strategic interactions in many economically relevant situations. For instance, noise and signals are often assumed to be distributed over a connected subset of the real numbers, which renders their support to be an infinite set. Similarly, players are often allowed to choose from an infinite set of pure strategies as, for example, in price or quantity competition, rank-order tournaments, and auctions.

Second, we provide a sufficient condition that allows deriving equilibrium existence in games between behavioral players with only one self. A large body of literature documents that people hold misspecified beliefs as they update in a non-Bayesian way (Enke and Zimmermann, 2017; Benjamin, Bodoh-Creed, and Rabin, 2019), they are overconfident (de la Rosa, 2011) or misperceive the mechanism through which signals are generated (Rabin, 2002; Rabin and Vayanos, 2010). How-

1. For an excellent overview see Eyster (2019).

ever, when analyzing strategic interactions between players, such deviations from standard rationality assumptions render equilibrium existence unclear. Our sufficient condition essentially requires that players' (incorrect) beliefs about the information structure and the strategy profile can be represented by a continuous mapping of the respective actual objects. As this condition is easily tested, it provides a ready to use toolbox for deriving equilibrium existence when studying the consequences of distorted beliefs or suboptimal decision making in strategic interaction.

To fix thoughts and illustrate our notation, Section 4.2 provides an example of a Dual-Selves game. We present the general model setup and state our main result in Section 4.3. Section 4.4, discusses how our existence result relates and contributes to Behavioral Game Theory. Section 4.5 concludes.

4.2 An Example of a Dual-Selves Game

This section works out a simple example of a Dual-Selves game that serves as an antidote to the abstractness of the following section, in which we define a general class of games between players with dual-selves. We consider a tournament in which each player consists of an *action self* that seeks to win the tournament and a *belief self* that values only self-esteem. Action selves are heterogeneous in ability and receive an informative signal about their ability before choosing effort in the tournament. Belief selves can manipulate how their corresponding action self interprets the ability signal to trigger effort decisions that improve future self-esteem. Hence, the two selves' objectives reflect the intra-personal conflict between trying to win the tournament and securing self-esteem. Jointly, they determine the player's behavior in the tournament.

Formally, there are two players, $i \in \{1,2\}$, each of which consists of an action self i and a corresponding belief self $\hat{i}$. The action selves have ability, $a^i \in \{a^i_L, a^i_H\}$ and choose effort, $x^i \in [0,1]$ at costs $c(x^i)$. With equal and independent probability, each action self is either of high ability, a_H, or low ability, a_L. For each action self, the tournament results in consequence, $y^i \in \{0,1\}$, where i wins the tournament if and only if $y^i = 1$. The consequence is determined by the action selves' abilities

and effort choices. Whenever an action self is of higher ability than his opponent, he wins the tournament. If abilities are identical, the action self with the higher effort wins the tournament. In the case of a tie in ability and effort, the realization of random variable z with $z \in \{1,2\}$ determines the winner. The state of the world $\omega \in \Omega$ is given by the realization of abilities and z such that $\omega = (a^i, a^j, z)$. To formalize the described relationship, we can construct a consequence function $f^i : \Omega \times [0,1] \times [0,1] \to \{0,1\}$ for each player that maps a state of the world and efforts into consequences. We assume that i receives benefit b if and only if he wins the tournament. As he also has to bear effort costs his payoff $\pi^i : [0,1] \times \{0,1\} \to \mathbb{R}$ depends on x^i and y^i.

Before choosing effort, each action self receives an informative signal, $s^i \in \{s_L, s_H\}$, about his ability. The signals' objective precision is given by $\mathbb{P}\left[a^i = a_H | s^i = s_H\right] = \mathbb{P}\left[a^i = a_L | s^i = s_L\right] = q^i$. However, the action self has a subjective belief about the signal's precision given by $\tilde{q}^i$, which is the choice variable of the belief self. Formally, $\hat{i}$ chooses $\tilde{q}^i \in [0,1]$ to maximize self-esteem, i.e., the action self's expected posterior probability to be of high ability. We denote this probability by $\mathbb{E}^i\left[\mathbb{P}^i\left[a_H | \tilde{q}^i, \tilde{q}^j, x^i, x^j, y^i\right]\right]$, where $j \neq i$ and the expectation is taken over all selves' strategy profile. The action self's subjective expected payoff is then given by

$$\mathbb{E}^i\left[\mathbb{P}^i\left[y^i = 1 | x^i, s^i, \tilde{q}^i\right] \cdot b - c\left(x^i\right)\right], \tag{4.1}$$

where $\mathbb{P}^i\left[y^i = 1 | x^i, s^i, \tilde{q}^i\right]$ denotes the subjective probability of winning the tournament, and the expectation is taken over the strategy profile of all selves.

Our paper will propose the Dual-Selves equilibrium as a solution concept for this type of strategic interaction with dual-selves. In a Dual-Selves equilibrium, action selves' effort choices maximize their expected utility given the subjective beliefs chosen by their respective belief self. Given these effort choices, the belief selves' distortions of the signal precision maximize each player's self-esteem.

This example is illustrative of the next sections in several ways. First, it shows how we can use standard notation to describe games with dual-selves. Second, it shows that in dual-selves games, beliefs and actions may influence each other in a non-trivial way. In particular, the players' self-esteem depends on the tournament's

outcome, his effort choice and how the belief self chooses the subjective belief about the signal structure of the game. In accordance, it is not immediately clear under what conditions an equilibrium in the dual self game exists.

4.3 A General Dual-Selves Game

This section describes a general model for games in which each player has a dual self—one self chooses actions, and one chooses beliefs. In order to capture a broad spectrum of objectives that belief selves may have, we explicitly model how actions and a state of the world determine consequences, which then determine payoffs.[2] We study games with the following timing. First, a state and a profile of signals are drawn. Second, selves may privately observe signals. Third, all selves simultaneously make their choices. Fourth, the profile of actions in combination with the state of the world determine consequences, which, in turn, determine payoffs.

Throughout the paper, we denote the action self of a player by i and the same player's belief self by $\hat{i}$. We assume that all selves interact non-cooperatively. A game between a finite number, N, of players whose behavior is defined by the decisions of her selves can be described by a tuple

$$\tilde{\mathscr{G}} = \langle \tilde{I}, \tilde{\Omega}, \tilde{\mathbb{S}}, \tilde{p}, \tilde{\mathbb{X}}, \tilde{\mathbb{Y}}, \tilde{f}, \tilde{\pi} \rangle,$$

with $\tilde{I}$ as the finite set of selves. Let $I = \{1, 2, \ldots, N\}$ denote the set of *action selves* and $\hat{I} = \{\hat{1}, \hat{2}, \ldots, \hat{N}\}$ the set of *belief selves* with $\tilde{I} = I \cup \hat{I}$. The complete and separable metric space $\tilde{\Omega}$ is the set of payoff-relevant states. Further, all action selves may privately observe a signal about the state of the world before taking actions. We collect the profiles of possible signals in the space $\tilde{\mathbb{S}} = \bigtimes_{i \in I} \mathbb{S}^i$, where $\mathbb{S}^i$ is the set of signals of self i. For all i, $\mathbb{S}^i$ is a complete and separable metric space.

2. Parts of this notation are borrowed from Esponda and Pouzo (2016). It allows us also to study situations in which belief selves do not only wish to manipulate beliefs about expected payoffs but also care about how payoffs come about, i.e., through which combination of strategies and state of the world. Such objectives may, for example, be relevant if belief selves try to anticipate a correct distribution over consequences, in line with the ideas in Esponda and Pouzo (2016).

The signal structure is represented by a probability measure $\tilde{p}$ over the Borel subsets[3] of $\tilde{\Omega} \times \tilde{\mathbb{S}}$. For simplicity, it is assumed to have marginals over $\tilde{\Omega}$ and $\tilde{\mathbb{S}}$ with full support; $p_{\tilde{\Omega}|S^i}\left(\cdot|s^i\right)$ denotes a probability measure over the subsets of $\tilde{\Omega}$ given $S^i = s^i$, and p_{S^i} denotes the marginal probability measure over the subsets of $\mathbb{S}^i$.

The action set of action self i is denoted by a compact metric space $\mathbb{X}^i$. Hence, $\mathbb{X} = \bigtimes_{i \in I} \mathbb{X}^i$ is the set of action profiles of all action selves. Each belief self chooses at least a part of the beliefs that the corresponding action self holds about the game's parameters or the strategies of her opponents. We denote the set of action profiles of the belief selves by $\hat{\mathbb{X}} = \bigtimes_{\hat{i} \in \hat{I}} \mathbb{X}^{\hat{i}}$, where $\mathbb{X}^{\hat{i}}$ is a compact metric space and contains the actions of belief self $\hat{i}$. Intuitively, the set $\mathbb{X}^{\hat{i}}$ thus contains those beliefs about the game of action self i, which can be influenced by belief self $\hat{i}$. If belief selves can manipulate an action self's belief about his opponents' strategies, this set would contain the strategy set of all other action selves. As in the example from Section 4.2, this set can, however, also include a collection of possible signal precisions. $\tilde{\mathbb{X}} = \mathbb{X} \times \hat{\mathbb{X}}$ is then the set of action profiles of all selves in the game. As Milgrom and Weber (1985), we allow all selves to play distributional strategies. A strategy of action self i is a probability measure σ^i on the subsets of $\mathbb{S}^i \times \mathbb{X}^i$, for which the marginal distribution on $\mathbb{S}^i$ is p_{S^i}. Formally, this restriction on the marginal distribution is that for all $T \subset \mathbb{S}^i$, $\sigma^i\left(T \times \mathbb{X}^i\right) = p_{S^i}(T)$. Moreover, the behavioral strategy of action self i is the conditional distribution denoted by $\sigma\left(\mathrm{d}x^i|s^i\right)$.

Let Σ^i denote the space of distributional strategies of action self i and $\Sigma = \bigtimes_{i \in I} \Sigma^i$ the space of all strategy profiles of the action selves. A strategy profile of all action selves is a vector of strategies $\sigma = \left(\sigma^i\right)_{i \in I} \in \Sigma$. For the belief selves, a strategy of $\hat{i}$ is a probability measure $\mu^{\hat{i}} \in \mathscr{M}\left(\mathbb{X}^{\hat{i}}\right)$ over $\mathbb{X}^{\hat{i}}$. A strategy profile of all belief selves is a vector of strategies $\mu = \left(\mu^{\hat{i}}\right)_{\hat{i} \in \hat{I}}$.

Moreover, let the finite set $\mathbb{Y}^j$ denote the set of (observable) consequences for self $j \in \tilde{I}$, such that $\mathbb{Y} = \bigtimes_{i \in I} \mathbb{Y}^i$ is the set of consequences for all action selves and $\hat{\mathbb{Y}} = \bigtimes_{i \in \hat{I}} \mathbb{Y}^i$ the set of consequences for all belief selves. Accordingly, $\tilde{\mathbb{Y}} = \mathbb{Y} \times \hat{\mathbb{Y}}$.

3. As long as not stated otherwise we will only consider Borel subsets for the rest of the paper.

For action self i, the consequence function $f^i : \mathbb{X}^i \times \mathbb{X}^{\hat{i}} \times \Omega \to \mathbb{Y}^i$ assigns a consequence to every combination of own action, the action of the associated belief self and the realized state of the world.[4] Note that action self i's realized consequence depends on the choice of her belief self $\hat{i}$, but does not directly depend on the other action selves' choices. More specifically, which consequence the action self anticipates does not immediately follow from the true strategy profile that is being played in equilibrium but rather from the beliefs about that strategy profile, which, in turn, are chosen by the belief self. This feature allows us to represent applications in which a player disregards or at least mis-predicts the impact of other players' actions on consequences.

For a given action and signal of action self i, each action $x^{\hat{i}} \in \mathbb{X}^{\hat{i}}$ of the belief self induces a distribution over consequences $Q^i_{x^{\hat{i}}}\left(\cdot|s^i, x^i\right)$, which is assumed to be continuous on $\mathbb{X}^{\hat{i}} \times \mathbb{S}^i \times \mathbb{X}^i$. Note that we allow the belief self $\hat{i}$'s choice to affect the perceived distribution over outcomes directly. Accordingly, belief self $\hat{i}$'s choice can not only distort action self i's perception about which actions are being played, but it might as well affect the likelihood with which action self i expects a given consequence to occur given all opponents' actions. Such a situation arises, for example, in the set-up described in Section 4.2, where the belief self manipulates the action self's perception of the signal's precision. The bounded payoff function is given by $\pi^i : \mathbb{X}^i \times \mathbb{X}^{\hat{i}} \times \mathbb{Y}^i \to \mathbb{R}$, which is assumed to be continuous on $\mathbb{X}^i \times \mathbb{X}^{\hat{i}}$. Here, the dependence of π^i on $x^{\hat{i}}$ allows to capture concepts in which beliefs directly enter the utility functions of players.

We allow the utilities of each belief self to depend on his own action and the strategies of all other action players. For example, an objective of a belief self may be to correctly anticipate the strategies of all action players or a distorted functional of these strategies (see Section 4.4 for details). Formally, each belief self $\hat{i}$ has the same consequence function given by $f^{\hat{i}} : \Sigma \to \Sigma$, with $f^{\hat{i}}(\sigma) = \sigma$. Hence, the observable

4. In contrast to the consequence function in the previous section, the action selves' general consequence functions do not directly depend on the actions of other action selves, but on each action selves' beliefs. In our example, the consequence function is a special case in which we require beliefs about other action selves actions to be correct.

consequence for the belief player is the strategy profile that is being played by all action selves. The payoff function of $\hat{i}$ is given by $\pi^{\hat{i}} : \Sigma \times \mathbb{X}^{\hat{i}} \to \mathbb{R}$ with $\pi^{\hat{i}}\left(\sigma, x^{\hat{i}}\right) \leq 0$.

The described utility functions allow us to derive under which conditions a strategy is optimal for an action self and a belief self. A strategy of a belief self, $\mu^{\hat{i}}$ is optimal if and only if $\mu^{\hat{i}}\left(x^{\hat{i}}\right) > 0$ implies

$$x^{\hat{i}} \in \arg\max_{\bar{x}^{\hat{i}} \in X^{\hat{i}}} \pi^{\hat{i}}\left(\sigma, \bar{x}^{\hat{i}}\right).$$

Similarly, a strategy σ^i of action self i is optimal if and only if $\sigma^i\left(x^i, s^i\right) > 0$ implies

$$x^i \in \arg\max_{\bar{x}^i \in \mathbb{X}^i} E_{\bar{Q}^i_{\mu^{\hat{i}}}(\cdot|s^i,\bar{x}^i)}\left[\pi^i\left(\bar{x}^i, x^{\hat{i}}, y^i\right)\right],$$

where $\bar{Q}^i_{\mu^{\hat{i}}}\left(\cdot|s^i, x^i\right) = \int_{\mathbb{X}^{\hat{i}}} Q^i_{x^i}\left(\cdot|s^i, x^i\right) \mu^{\hat{i}}\left(\mathrm{d}x^{\hat{i}}\right)$ is the distribution over consequences for action self i, conditional on $\left(s^i, x^i\right) \in \mathbb{S}^i \times \mathbb{X}^i$, induced by the strategy of the belief self $\mu^{\hat{i}}$. We define a Dual-Selves equilibrium as follows:

Definition 4.1. *A strategy profile $\left(\sigma^*, \mu^*\right)$ is a Dual-Selves equilibrium of game $\tilde{\mathcal{G}}$ if and only if, (i) σ^{i*} is optimal given $\mu^{\hat{i}*}$ $\forall$ i, and (ii) $\mu^{\hat{i}*}$ is optimal given σ^* $\forall$ $\hat{i}$.*

Our definition embraces the notion that in decision processes with dual-selves, each self optimizes his actions taken the other selves' behavior as given. This approach seems in line with evidence on the brain's functioning when resolving intrapersonal conflict (Brocas and Carrillo, 2008; Alonso, Brocas, and Carrillo, 2013). However, it also puts clear boundaries on which kind of processes we capture in our analysis. In particular, we exclude decision processes in which actions immediately affect beliefs. In such processes, individuals choose actions while being aware that these actions will immediately also change their beliefs about future outcomes (see for example Kőszegi and Rabin, 2007; Masatlioglu and Raymond, 2016).

The definition of a Dual-Selves equilibrium presumes that all action selves best respond to their beliefs and all belief selves play a best response to the equilibrium strategy profile. Note that the definition does not directly require all action selves to play mutually best responses. In particular, each belief self chooses what the corresponding action self believes about the strategy profile that is being played.

Therefore, action selves optimize their action, given their beliefs rather than the true equilibrium strategy profile.

However, given the definition of a Dual-Selves equilibrium, action selves best-respond to their beliefs about other action selves' strategies. Hence, they best-respond to the other action selves' actual strategies whenever belief selves choose the correct belief about these strategies on the equilibrium path. Consequently, the definition can also be viewed as a generalization of Nash equilibrium for games in which every player only has one self. Nevertheless, we do not capture all possible Dual-Selves games. For example, we do not allow belief selves to condition beliefs on the signals that action selves receive. We will prove the following claim:

Theorem 4.1. *Suppose that* $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu^{\hat{i}}(\mathrm{d}x^{\hat{i}})$ *is continuous on* $\mathscr{M}(\mathbb{X}^{\hat{i}}) \times \Sigma \ \forall \ \hat{i}$. *Then there exists a Dual-Selves equilibrium in* $\tilde{\mathscr{G}}$.

We relegate the proof of the theorem to the appendix. In parts, the proof follows the arguments in Milgrom and Weber (1985). However, in Dual-Selves games, players do not interact immediately with each other but rather through the belief selves' choices. Therefore, we do not have to ensure continuity of all players' expected utilities in their opponents' strategies. Instead, we show under what conditions the action selves' expected utility is continuous in the corresponding belief self's strategy.

We consider the assumptions we have imposed on the action selves' action sets and the signal generating mechanism as relatively innocuous as they resemble conditions that are often invoked to ensure the existence of a Nash-Equilibrium in Single-Self games (see also Milgrom and Weber, 1985). Instead, the crucial additional requirement that Dual-Selves games have to fulfill to ensure the existence of a Dual-Selves equilibrium is the continuity of the belief selves' objective functions. As the applications of Theorem 4.1 in Section 4.4 in the appendix show, these requirements restrict the classes of games captured by our existence result. However, the theorem also shows for which kind of Dual-Selves games equilibrium existence is less problematic: If, for example, the set of possible beliefs $\hat{\mathbb{X}}$ is finite or the objectives of belief selves is to match one particular belief according to a continuous scoring rule, the condition in Theorem 4.1 is ensured. While the theorem explicitly addresses

situations with many players, it is also directly applicable to an individual-decision problem with dual-selves. In this case, the decision problem is only well defined if an equilibrium between the two selves exists for which the continuity restriction is essential.

4.4 Relation to Behavioral Game Theory

This section illustrates how our setup and results contribute to the literature that studies games in which each player only has one self but may have misspecified beliefs or belief dependent utility. First, we show how Dual-Selves equilibria correspond to equilibria in Single-Self games with not fully rational players. Second, we discuss several existence results in the literature that we extend to non-finite games as corollaries of Theorem 4.1. Third, we derive a simple sufficient condition that can be applied to test equilibrium existence if behavioral players interact strategically.

4.4.1 Relation to Other Equilibrium Concepts

Theorem 4.1 proves the existence of Dual-Selves equilibria. Notably, many equilibrium concepts in Behavioral Game Theory that feature behavioral players with only one self are equivalent to a Dual-Selves equilibrium. To delineate this correspondence, consider the action space $\mathbb{X}^{\hat{i}}$ and the objective function $\pi^{\hat{i}}(\cdot)$ of the belief self. Essentially, $\mathbb{X}^{\hat{i}}$ determines *which* parts of the game the behavioral player may misperceive, whereas the specification of $\pi^{\hat{i}}(\cdot)$ determines *how* these parts are misperceived.

To clarify this point, we will start this section by showing how one specific equilibrium concept – Cursed Equilibrium by Eyster and Rabin (2005) – can be accommodated as a Dual-Selves equilibrium. Recall that cursed equilibrium assumes that each player correctly predicts all opponents' actions but may underestimate to what extent these actions are correlated with the privately observed signals. We will first use our notation to describe a game in which every player has only one self. In a second step, we then show the analogy.

For the first step, consider game $\tilde{\mathcal{G}}$ defined in Section 4.3. Let game $\mathcal{G}_{OS}$ be equivalent to game $\tilde{\mathcal{G}}$ except for the following two changes. First, assume that the set of belief selves $\hat{I}$ is empty. Second, adapt the consequence function of action self i such that it is given by $f^i : \mathbb{X} \times \Omega \to \mathbb{Y}^i$. Hence, which consequence materializes depends on the state of the world and the actions of all opponents. Game $\mathcal{G}_{OS}$ is, therefore, a standard game in which each player has only one self (see also Esponda and Pouzo, 2016).

According to Eyster and Rabin (2005) a Cursed Equilibrium of $\mathcal{G}_{OS}$ can be defined as follows. Let σ^{-i} be the strategy profile of all opponents' strategies. Then a cursed player expects with probability $\mathcal{X} \in [0,1]$ that his opponents play

$$\bar{\sigma}^{-i}(x^{-i}|s^i) = \int_{\mathbb{S}^{-i}} \prod_{j \neq i} \sigma^j \left(\mathrm{d}x^j | s^j\right) \times \int_{\Omega} p_{\Omega \times \mathbb{S}^{-i}|\mathbb{S}^i}(\mathrm{d}\omega, \mathrm{d}s^{-i}|s^i).$$

For player i of type s^i, $\bar{\sigma}^{-i}(x^{-i}|s^i)$ denotes the average strategy of his opponents, where the average is taken over their types. The perceived probability distribution over outcomes of a cursed player might, therefore, be incorrect and is given by:

$$Q^i_{\bar{\sigma}}\left(y^i|s^i,x^i\right) = \int_{\Omega \times \mathbb{X}^{-i} \times \mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}} \Big[\mathcal{X}\, \bar{\sigma}^{-i}\left(\mathrm{d}x^{-i}|s^i\right) + (1-\mathcal{X}) \prod_{j \neq i} \sigma^{-i}\left(\mathrm{d}x^{-i}|s^{-i}\right) \Big] p_{\Omega \times S^{-i}|S^i}\left(\mathrm{d}\omega, \mathrm{d}s^{-i}|s^i\right).$$

A *cursed equilibrium* of $\mathcal{G}_{OS}$ is defined as follows:

Definition 4.2. *A strategy profile σ is a Cursed Equilibrium of game $\mathcal{G}_{OS}$ if, for all players $i \in I$ and $s^i \in \mathbb{S}^i$, and any x^i with $\sigma^i(x^i|s^i) > 0$, we have*

$$x^i \in argmax_{x^i \in \mathbb{X}^i} E_{Q^i_{\bar{\sigma}}(\cdot|s^i,x^i)}[\pi(x^i,y^i)].$$

To show that this concept is equivalent to a Dual-Selves equilibrium, consider game $\tilde{\mathcal{G}}_{OS}$. This game is identical to game $\mathcal{G}_{OS}$ except for three changes. First, for every action self i, there exists a belief self $\hat{i}$. Second, the consequence that self i anticipates depends on the action of his belief self instead of the strategies of his opponents. However, the action space of the belief self is given by Σ. Essentially,

the belief self, thus, selects what the player thinks about the strategy profile that is being played. More specifically, the consequence function is specified exactly as before but maps from the belief self's choice and the state of the world to outcomes, $f^i : \mathbb{X}^{\hat{i}} \times \Omega \to \mathbb{Y}^i$. Third, let the objective of the belief self be given by:[5]

$$\pi^{\hat{i}}(\sigma, x^{\hat{i}}) = -\sum_{y^i \in Y^i} (Q^i_{\sigma}(y^i|x^i, s^i) - Q^i_{x^{\hat{i}}}(y^i|x^i, s^i))^2.$$

By Theorem 4.1, we know that a Dual-Selves equilibrium in this game exists, whenever $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}}) \mu^{\hat{i}}(\mathrm{d}x^{\hat{i}})$ is continuous. Corollary 4.2, which we relegate to the appendix, derives sufficient conditions for this to be the case. In this Dual-Selves equilibrium, every belief self will play a best response to all action selves and all other belief selves. In equilibrium, each action self will, therefore, anticipate exactly the probability distribution over consequences that a cursed player would anticipate (see also Equation (4.A.2)). Accordingly, action self i will play a best response given that his beliefs are cursed (cp. Equation (4.A.1)). Hence, the strategy profile played by the action selves in the Dual-Selves equilibrium of $\tilde{\mathcal{G}}_{OS}$ constitutes a *cursed equilibrium* in $\mathcal{G}_{OS}$. We conclude that the concept of cursed equilibrium is identical to a Dual-Selves equilibrium with specific objective functions of the belief selves.

As implied by Definition 4.1, an important requirement to interpret a given equilibrium concept as a Dual-selves equilibrium is that actions and beliefs of a player can be considered to arise as a resolution of intra-personal conflict. Nevertheless, the correspondence between Dual-Selves equilibria and different equilibrium concepts in games with behavioral players is not restricted to Cursed Equilibrium. Many existing concepts can be reinterpreted in this way. There are four main classes of concepts that we can grasp. First, concepts in which behavioral payers systematically misperceive their opponents' strategies as, for example, in Jehiel (2005) and Eyster and Rabin (2005). Our setup allows us to capture such concepts by letting the

5. This objective function of the belief player illustrates that in the context of equilibrium concepts with misspecified beliefs, the belief player is to be interpreted as an as if construct, i.e., players behave in equilibrium as if a third party chose their beliefs according to the specified utility function. In other domains, the interpretation of dual-selves can be more explicit. For example, in the realm of image concerns, it makes sense to interpret the two selves indeed as parts of the individual's psychology.

belief self choose the expectation about the opponent's strategies. Hence, $\mathbb{X}^{\hat{i}} = \Sigma^{-i}$. Via an appropriate payoff function $\pi^{\hat{i}}$, the belief self is then incentivized to select the biased belief specified by the specific equilibrium concept.

Second, Dual-Selves equilibria also subsume concepts in which behavioral players hold fully rational beliefs in equilibrium, but their utility directly depends on beliefs as, for example, in Geanakoplos, Pearce, and Stacchetti (1989) or Dato et al. (2017). If $\pi^{\hat{i}}(\cdot)$ is specified such that $\sigma \in \mathbb{X}^{\hat{i}}(\sigma)$ holds in equilibrium, the Dual-Selves equilibrium choices of belief selves will, in fact, imply that each action self behaves as if he correctly anticipates the strategies played in equilibrium. As payoffs π^i are, however, allowed to depend on beliefs in our setup, we can allow for the utility to be belief dependent (For a formal derivation, see Corollaries 4.3 and 4.4).

Third, our setup can incorporate concepts in which players correctly anticipate their opponents' behavior but may hold a wrong view about the fundamentals of the game. Such biases are prominently discussed and well documented in the domain of individual decision-making. One possible source for such misperceptions is if signals are processed in a non-Bayesian way (see for examples Phillips and Edwards, 1966; Kahneman and Tversky, 1973; Rabin and Schrag, 1999; Bénabou and Tirole, 2011; Benjamin, Rabin, and Raymond, 2016), or individuals believe in a wrong signal generating mechanism (see for examples Barberis, Shleifer, and Vishny, 1998; Rabin, 2002; Rabin and Vayanos, 2010). Strategic interaction between behavioral players that fall victim to such belief biases can be studied by using Dual-Selves equilibria. To capture such biases, each belief self not only chooses a player's belief about his opponent's strategies but also about primitives of the game (such as the signal generating mechanism as in our example from Section 4.2). This can easily be incorporated in the action set of belief selves, which is only restricted to be a metric space. Again, $\pi^{\hat{i}}(\cdot)$ then specifies according to which rule a belief self chooses the corresponding beliefs.

Fourth, we can also capture concepts that allow for misperceptions concerning both the strategy profile played in equilibrium and primitives of the game. Embracing the notion that both of these components may be misperceived typically implies

that beliefs about the game's strategies and primitives are jointly determined in equilibrium according to some optimality criterion as in Esponda and Pouzo (2016) or a consistency criterion as in Esponda (2008). This optimality criterion then corresponds to the belief selves' objectives (see also the proof of Corollary 4.5).

4.4.2 Generalization of Equilibrium Concepts

As argued in the previous section, various existing equilibrium concepts correspond to Dual-Self equilibria. Corollaries 4.2 – 4.5 in the appendix show that Cursed Equilibrium by Eyster and Rabin (2005), Personal Nash Equilibrium by Dato et al. (2017), Psychological Equilibrium by Geanakoplos, Pearce, and Stacchetti (1989), and Berk-Nash Equilibrium by Esponda and Pouzo (2016), for example can be reinterpreted as Dual-Selves equilibria. A key virtue of Theorem 4.1 is that it derives equilibrium existence also for non-finite games. As all of the above contributions restrict themselves to show equilibrium existence in games with finite action and signal spaces, Theorem 4.1 together with Corollaries 4.2 – 4.5, show how the corresponding equilibrium existence generalizes to non-finite games.

4.4.3 Sufficient Condition for Equilibrium Existence

When modeling individual decisions, many papers in behavioral economics have deviated from standard rationality assumptions that are also a necessary prerequisites to apply the prominent equilibrium existence results in game theory. This section derives a sufficient condition to extend these results to equilibrium concepts that feature behavioral players.

Suppose that there is some equilibrium concept $\mathscr{C}$ in which players are behavioral and a corresponding game $\mathscr{G}_{OS}$ in which each player has one self (see also Section 4.4.1). Furthermore, collect the strategies and the distribution over types in the set $\Theta = \Sigma \times \mathscr{M}(\Omega \times S)$. Given the strategy profile σ and the true signal structure $\tilde{p}$, equilibrium concept $\mathscr{C}$ describes for each player i which (incorrect) beliefs he holds about these two objects. We assume that this process can be represented by a function $F^i_{\mathscr{C}} : \Theta \to \Theta$, which maps the true $\theta \in \Theta$ in a vector $\hat{\theta}^i \in \Theta$ which describes

what i believes.[6] More specifically, player i anticipates the distribution over outcomes to be identical to $Q^i_{\hat{\theta}^i}\left(y^i|s^i,x^i\right)$ (cp. Section 4.3). We define a $\mathscr{C}$-equilibrium in game $\mathscr{G}_{OS}$ as follows:

Definition 4.3. *A $\mathscr{C}$-equilibrium in $\mathscr{G}_{OS}$ is a strategy profile σ such that:*

(i) Every player i plays a best response given $\hat{\theta}^i$.

(ii) If $\mathscr{G}_{OS}$ specifies $\tilde{p}$ and the players play strategy profile σ, then each player i holds the belief $\hat{\theta}^i = F^i_{\mathscr{C}}(\sigma,\tilde{p})$.

To illustrate what kind of concepts this specification can grasp, note two points. First, recall that space Σ is the space of profiles of distributional strategies. Hence, a misperception about opponents' strategies can be represented by some collection of $F^i_{\mathscr{C}}$'s by assuming that the belief about the profile of distributional strategies is distorted. Second, other concepts may assume that players misperceive their own or other players' types or other parts of the information structure, which can be reflected if $F^i_{\mathscr{C}}$ specifies that a player holds a wrong belief about $\tilde{p}$.

Corollary 4.1. *If $F^i_{\mathscr{C}}(\cdot)$ is continuous on Θ $\forall$ i, there exists a $\mathscr{C}$-equilibrium in $\mathscr{G}_{OS}$.*

Due to its simplicity, the condition in Corollary 4.1 lends itself to be applied in various contexts if behavioral players interact strategically and equilibrium existence is not obvious. There are two common subcases of Corollary 4.1. First, equilibrium existence should rarely be an issue if all players have belief-dependent utility but correctly anticipate the distribution of types and the strategy profile. Such situations arise for example if players have image concerns as in Bénabou and Tirole (2006b) or anticipatory utility as in Caplin and Leahy (2001). In our notation, these concepts correspond to $F^i_{\mathscr{C}}$ being the identity for all players.[7] Similarly, how

6. As an easy example consider a player who receives a correct signal about his ability $\theta \in [0,1]$. However, the player is overconfident and believes to have ability $\hat{\theta} = f(\theta)$, where $f : [0,1] \to [0,1]$ is a continuous function. Then his perceived ability is clearly a continuous function of the underlying true ability.

7. Notably, our concept covers additional cases compared to psychological games as discussed in Geanakoplos, Pearce, and Stacchetti (1989) even if we restrict $F^i_{\mathscr{C}}$ to be the identity for all i. In particular, we can allow for belief dependent utility also for own strategies, which may arise if players are expectation based loss averse as in Dato et al. (2017) or with respect to their type, which may, for example, arise if players have image concerns with respect to their ability.

we oftentimes model non-Bayesian updating such as correlation neglect (Enke and Zimmermann, 2017), base rate neglect (Benjamin, Bodoh-Creed, and Rabin, 2019) or overconfidence (de la Rosa, 2011) relies on formulations in which individuals' beliefs correspond to a continuous distortion of their true signal or type. In these cases, it is obvious that also the corresponding $F^i_{\mathscr{C}}$ is going to be continuous, and a $\mathscr{C}$-equilibrium exist if players fall victim to such belief distortions in strategic situations.

While a continuous $F^i_{\mathscr{C}}$ can grasp many cases in the literature, the assumptions on $F^i_{\mathscr{C}}$ are not innocuous. First, we assume that the function maps into Θ. This excludes every concept in which players expect types or strategies that, in fact, do not exist in the game. Second, we impose continuity of $F^i_{\mathscr{C}}$, which for example, excludes concepts in which players may only be partially attentive in the sense that they only realize that specific actions are played in equilibrium if their likelihood of being on the equilibrium path exceeds some threshold (Gabaix, 2012, 2014).

4.5 Conclusion

This paper conceptualizes the strategic interaction between players whose beliefs and actions follow from different or even opposing objectives and do not necessarily constitute coherent consequences of each other, i.e., players with dual-selves. To account for the fact that the players themselves and the dual selves of each player interact non-cooperatively, we have defined the solution concept Dual-Selves equilibrium, which can be viewed as a generalization of Nash equilibrium to study Dual-Selves games. Our main theorem shows that the continuity of the belief selves' objective functions ensures a Dual-Selves equilibrium.

While we explicitly allow (i) multiple players and (ii) the dual-selves of each player to interact at the same time, our results also provide useful insights to the analysis of settings in which there is only one individual with a dual self or in which there are multiple players with single-selves but cognitive limitations. First, Dual-Selves equilibrium can be applied to study decisions that arise as a resolution to the intra-personal conflict in individual decision-making. Even these decisions are inher-

ently strategic due to the opposing objectives of the selves. Second, we show that a variety of solution concepts proposed in Behavioral Game Theory, where players have only one self but are characterized by limited cognitive capabilities or non-standard preferences (e.g., Geanakoplos, Pearce, and Stacchetti, 1989; Eyster and Rabin, 2005; Esponda and Pouzo, 2016; Dato et al., 2017), constitute subcases of Dual-Selves equilibrium. By proving the existence of Dual-Selves equilibrium in non-finite games, we extend the previously established existence results, which are restricted to finite games. We further contribute to this strand of the literature by providing a simple sufficient condition that allows deriving equilibrium existence in games between behavioral players with only one self.

By formalizing the notion of a Dual-Selves game and a Dual-Selves equilibrium, we provide a well-defined framework to study the decisions of individuals with dual selves. We hope that this framework also motivates future empirical research. In various setups, there is mounting evidence that the resolution of intra-personal conflict governs choices. Applying our framework will allow researchers to derive additional theoretical hypotheses and to delineate situations in which dual-selves processes are an essential driver of economic choices from situations in which it is sufficient to think of decision-makers as one entity.

Appendix 4.A Proofs

Proof of Theorem 4.1. To prove the claim, we show that $\mathscr{G}$ satisfies the requirements of the existence result by Glicksberg (1952).[8]

Step 1: Compactness and Convexity of Strategy Spaces.

(a): Action Selves. We will prove the existence of a Dual-Selves equilibrium in distributional strategies in $\mathscr{G}$. For this purpose, we start by arguing that the strategy spaces are compact and convex metric spaces in the weak topology. Note that $\tilde{p}$ is tight since it is a measure over complete separable metric spaces (see Parthasarathy (1967), Theorem 3.2, p.29). Additionally, action spaces are assumed to be compact, and, hence, they are also complete (see Theorem 8.16, p.96, in De la Fuente (2000)) such that also the Cartesian products $\mathbb{X}^i \times \mathbb{S}^i$ are complete. By Theorem 3.2 in Parthasarathy (1967), each action self's set of distributional strategies is then a tight set of probability measures[9]; also, since it is a set of measures over the Cartesian product of separable and complete metric spaces[10], it is complete itself by Theorem 6.5 in Parthasarathy (1967) and, hence, closed in the weak topology. By Prohorov's Theorem[11], it follows that the strategy sets are compact metric spaces in the weak topology. Furthermore, the set of distributional strategies is clearly convex.

(b): Belief Selves: By the same arguments as in *Step 1a*, the belief selves' strategy sets are compact metric spaces in the weak topology and the set of strategies is clearly convex.

Step 2: Continuity of the action selves' utility. Note, that the expected utility of an action self i is given by

$$U^i\left(\mu^{\hat{i}}, \sigma^i\right) = \int E_{\bar{Q}^i_{\mu^{\hat{i}}}(\cdot|s^i,x^i)}\left[\pi^i\left(x^i, x^{\hat{i}}, y^i\right)\right]\sigma^i\left(\mathrm{d}x^i|s^i\right)p_{\mathbb{S}^i}\left(\mathrm{d}s^i\right)$$

8. Specifically, we will use the following result: Let the players' strategy spaces be nonempty compact, convex subsets of convex Hausdorff linear topological spaces. Let the payoff functions be continuous on the product of the strategy spaces, and let each player's payoff function be quasiconcave in his strategy. Then an equilibrium point exists.

9. A set of probability measures on a metric space is called tight if for every $\epsilon > 0$ there is a compact set K such that for every P in the set of measures, $P(K) > 1 - \epsilon$.

10. The Cartesian product of separable and complete metric spaces is separable itself (see De la Fuente, 2000, p. 82)

11. Billingsley (1968), Theorem 6, p.240.

$$= \int_{\mathbb{X}^i \times \mathbb{S}^i \times \mathbb{X}^{\hat{i}}} \sum_{y^i \in \mathbb{Y}^i} \pi^i\left(x^i, x^{\hat{i}}, y^i\right) Q^i_{x^{\hat{i}}}\left(y^i | s^i, x^i\right) \mathrm{d}\mu^{\hat{i}}\left(x^{\hat{i}}\right) \mathrm{d}\sigma^i\left(x^i, s^i\right),$$

where we make use of Fubini's theorem. To prove continuity under the weak topology, we need to prove the following: For any sequence $\left(\mu^{\hat{i}}_n, \sigma^i_n\right)_n$ such that $\left(\mu^{\hat{i}}_n, \sigma^i_n\right) \Rightarrow \left(\mu^{\hat{i}}, \sigma^i\right)$ it has to hold that $\lim_{n\to\infty} U^i\left(\mu^{\hat{i}}_n, \sigma^i_n\right) = U^i\left(\mu^{\hat{i}}, \sigma^i\right)$. Define the probability measure $p^i_n\left(x^i, s^i, x^{\hat{i}}\right)$ such that $\int g \mathrm{d}\mu^{\hat{i}}_n\left(x^{\hat{i}}\right) \mathrm{d}\sigma^i_n\left(x^i, s^i\right) = \int g dp^i_n\left(x^i, s^i, x^{\hat{i}}\right)$. By Theorem 1.1 in Feinberg, Kasyanov, and Zadoianchuk (2014), $\lim_{n\to\infty} \int g dp^i_n\left(x^i, s^i, x^{\hat{i}}\right) = \int g dp^i\left(x^i, s^i, x^{\hat{i}}\right)$ for all bounded and continuous real valued functions g if $p^i_n \Rightarrow p^i$. Note that by the continuity and boundedness of payoffs and $Q_{x^{\hat{i}}}$ also $\sum_{y^i \in \mathbb{Y}^i} \pi^i\left(x^i, y^i, x^{\hat{i}}\right) Q^i_{x^{\hat{i}}}\left(y^i | s^i, x^i\right)$ is continuous and bounded implying that $\lim_{n\to\infty} U^i\left(\mu^{\hat{i}}_n, \sigma^i_n\right) = U^i\left(\mu^{\hat{i}}, \sigma^i\right)$ if $\left(\mu^{\hat{i}}_n, \sigma^i_n\right) \Rightarrow \left(\mu^{\hat{i}}, \sigma^i\right)$. Hence, U^i is continuous. For later reference also note that $\sum_{y^i \in \mathbb{Y}^i} \pi^i\left(x^i, y^i, x^{\hat{i}}\right) Q^i_{x^{\hat{i}}}\left(y^i | s^i, x^i\right)$ is bounded, such that $U^i\left(\mu^{\hat{i}}, \sigma^i\right)$ is an affine function of $\sigma^i\left(x^i, s^i\right)$.

Step 3: Existence of Nash-Equilibrium:

Since $U^{\hat{i}}\left(\sigma, \mu^{\hat{i}}\right)$ is an integral over a non-negative function, it is an affine function on the set of strategies of the belief self $\hat{i}$, and, hence, quasi-concave. In summary, the selves' strategy sets are compact, convex metric spaces and the payoff functions are continuous and affine. According to the version of Glicksberg's theorem that is also applied by Milgrom and Weber (1985) these properties guarantee the existence of a Bayes-Nash Equilibrium. In each Bayes' Nash Equilibrium, the selves play mutually best responses. Consequently $\sigma^i\left(x^i, s^i\right) > 0$ only if

$$x^i \in arg\max_{\bar{x}^i \in \mathbb{X}^i} E_{\tilde{Q}^i_{\mu^{\hat{i}}}(\cdot|s^i, \bar{x}^i)}\left[\pi^i\left(\bar{x}^i, x^{\hat{i}}, y^i\right)\right]. \tag{4.A.1}$$

Similarly, $\mu^{\hat{i}}\left(x^{\hat{i}}\right) > 0$ only if

$$x^{\hat{i}} \in \mathbb{X}^{\hat{i}}(\sigma) \equiv arg\max_{x^{\hat{i}} \in \mathbb{X}^{\hat{i}}} \pi^{\hat{i}}\left(\sigma, x^{\hat{i}}\right). \tag{4.A.2}$$

□

Proof of Corollary 4.1. Take game $\mathcal{G}_{OS}$ and transfer it to a game $\tilde{\mathcal{G}}_{OS}$ in which each player has two selves a long the lines laid out in Section 4.4.1. In particular, the

action self corresponds to the player in $\mathcal{G}_{OS}$. The belief self $\hat{i}$ receives no signal, has strategy space $\mathbb{X}^{\hat{i}} = \Theta$ and objective

$$\pi^{\hat{i}}(\sigma, x^{\hat{i}}) = -\sum_{y^i \in Y^i} \left[Q_{F^i_{\mathscr{C}}(\sigma, p_\Omega)}(y^i|x^i, s^i) - Q_{x^{\hat{i}}}(y^i|x^i, s^i) \right]^2.$$

As $\pi^{\hat{i}}$ is continuous due to the continuity of $F^i_{\mathscr{C}}$, according to Theorem 4.1, there exist a Dual-Selves equilibrium in $\tilde{\mathcal{G}}_{OS}$. This Dual-Selves equilibrium corresponds to a $\mathscr{C}$-equilibrium in $\mathcal{G}_{OS}$. □

Appendix 4.B Applications

In this section, we are going to highlight the usefulness of Theorem 1 by extending a number of existence results for finite games to non-finite games. In particular, we extend cursed equilibrium by Eyster and Rabin (2005) (Section 4.B.1), Personal Nash Equilibrium by Dato et al. (2017) (Section 4.B.2), Psychological Equilibrium by Geanakoplos, Pearce, and Stacchetti (1989) (Section 4.B.3), and Berk-Nash Equilibrium by Esponda and Pouzo (2016) (Section 4.B.4). The game in which each player only has one self $\mathcal{G}_{OS}$ constitutes the starting point for this analysis. In any Nash-Equilibrium of $\mathcal{G}_{OS}$, action selves form correct beliefs according to the true distribution over consequences, which is given by

$$Q_\sigma\left(y^i|s^i, x^i\right) = \int_{\Omega \times \mathbb{X}^{-i} \times \mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i, x^{-i}, \omega) = y^i\}} \prod_{j \neq i} \sigma^{-i}\left(\mathrm{d}x^{-i}|s^{-i}\right) \times p_{\Omega \times S^{-i}|S^i}\left(\mathrm{d}\omega, \mathrm{d}s^{-i}|s^i\right).$$

Following Esponda and Pouzo (2016) we refer to this distribution as the *objective* distribution over consequences throughout this section. Moreover, we make the following additional assumptions about game $\mathcal{G}_{OS}$:

Assumption 4.1. *For all $i \in I$, $p_{\Omega \times \mathbb{S}^{-i}|\mathbb{S}^i}$ is absolutely continuous with respect to the product measure $p_\Omega \times p_{\mathbb{S}^1} \times \ldots p_{\mathbb{S}^{i-1}} \times p_{\mathbb{S}^{i+1}} \cdots \times p_{\mathbb{S}^N}$.*

Assumption 4.1 is reminiscent of the first requirement in Milgrom and Weber (1985). However, it requires absolute continuity of *conditional* probability measures instead of *unconditional* probability measures. We consider this assumption to be a rather mild restriction on feasible distributions of states and signals: it is trivially

fulfilled if for all players (i) Ω and $\mathbb{S}^i$ are finite, (ii) states and the players' signals are independent, or (iii) $p_{\Omega\times\mathbb{S}^{-i}|\mathbb{S}^i}$ has no mass points (see also Milgrom and Weber, 1985). Assumption 4.1 allows us to simplify the expression for the objective distribution of outcomes such that

$$\begin{aligned} Q_\sigma\left(y^i|s^i,x^i\right) &= \int_{\Omega\times\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}} g_i\left(\omega,s^{-i}|s^i\right) p_\Omega\left(\mathrm{d}\omega\right) \prod_{j\neq i}^{N} \sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right) \\ &= \int_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathrm{Prop}\left(y^i|x,s\right) \prod_{j\neq i}^{N} \sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right), \end{aligned}$$

where $g_i\left(\omega,x^{-i}|s^i\right)$ is the Radon-Nikodym derivative. Furthermore, we assume this probability to be a continuous function.

Assumption 4.2. *Prop* $(\cdot|x,s)$ *is a continuous function on* $\mathbb{X}\times\mathbb{S}$.

Assumptions 4.1 and 4.2 together ensure that as a corollary of Theorem 4.1 there exist a Nash-Equilibrium in the game $\mathcal{G}_{OS}$, i.e. an equilibrium, when all players form correct beliefs.

4.B.1 Existence of Cursed Equilibrium

Consider a Cursed equilibrium as defined in Section 4.4.1. We make the following assumptions on the game $\mathcal{G}_{OS}$ to be able to transfer the notion of a Cursed Equilibrium to our setting, where $g_i\left(s^{-i}|s^i\right) = \int_\Omega g_i\left(\omega,x^{-i}|s^i\right) p_\Omega\left(\mathrm{d}\omega\right)$ and $g_i\left(\omega|s^i\right) = \int_{\mathbb{S}^{-i}} g_i\left(\omega,x^{-i}|s^i\right) p_{S^{-i}|S^i}(ds^{-i})$.

Assumption 4.3. $\int_\Omega \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}} g_i\left(s^{-i}|s^i\right) \times g_i\left(\omega|s^i\right) p_\Omega\left(\mathrm{d}\omega\right)$ *is bounded and continuous on* $\mathbb{X}\times\mathbb{S}$.

The following corollary extends the existence result in Eyster and Rabin (2005) to non-finite games.

Corollary 4.2. *Suppose that Assumptions 4.1–4.3 hold for game* $\mathcal{G}_{OS}$. *Then a Cursed Equilibrium exists.*

Proof. Take $Q^i_{\bar\sigma}\left(y^i|s^i,x^i\right)$ as defined in Section 4.4.1. First note that because of Assumptions 4.1–4.3, we can write

$$
\begin{aligned}
&Q^i_{\bar\sigma}\left(y^i|s^i,x^i\right)\\
&= \mathscr{X}\int\limits_{\Omega\times\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}\bar\sigma^{-i}\left(\mathrm{d}x^{-i}|s^i\right)\times p_{\Omega\times S^{-i}|S^i}\left(\mathrm{d}\omega,\mathrm{d}s^{-i}|s^i\right)\\
&\quad+(1-\mathscr{X})\int\limits_{\Omega\times\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}\prod_{j\neq i}\sigma^{-i}\left(\mathrm{d}x^{-i}|s^{-i}\right)\times p_{\Omega\times S^{-i}|S^i}\left(\mathrm{d}\omega,\mathrm{d}s^{-i}|s^i\right)\\
&= \mathscr{X}\int\limits_{\Omega\times\mathbb{X}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}\bar\sigma^{-i}\left(\mathrm{d}x^{-i}|s^i\right)\times\int\limits_{\mathbb{S}^{-i}} p_{\Omega\times S^{-i}|S^i}\left(\mathrm{d}\omega,\mathrm{d}s^{-i}|s^i\right)\\
&\quad+(1-\mathscr{X})\int\limits_{\Omega\times\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}\prod_{j\neq i}\sigma^{-i}\left(\mathrm{d}x^{-i}|s^{-i}\right)\times p_{\Omega\times S^{-i}|S^i}\left(\mathrm{d}\omega,\mathrm{d}s^{-i}|s^i\right)\\
&= \mathscr{X}\int\limits_{\Omega\times\mathbb{X}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}\bar\sigma^{-i}\left(\mathrm{d}x^{-i}|s^i\right)\times\int\limits_{\mathbb{S}^{-i}} g_i\left(\omega,s^{-i}|s^i\right)\prod_{j\neq i}p_{\mathbb{S}^j}\left(\mathrm{d}s^j\right)p_\Omega\left(\mathrm{d}\omega\right)\\
&\quad+(1-\mathscr{X})\int\limits_{\Omega\times\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}\prod_{j\neq i}\sigma^{-i}\left(\mathrm{d}x^{-i}|s^{-i}\right)\times p_{\Omega\times S^{-i}|S^i}\left(\mathrm{d}\omega,\mathrm{d}s^{-i}|s^i\right)\\
&= \mathscr{X}\int\limits_{\Omega\times\mathbb{X}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}\bar\sigma^{-i}\left(\mathrm{d}x^{-i}|s^i\right)\times g_i\left(\omega|s^i\right)p_\Omega\left(\mathrm{d}\omega\right)\\
&\quad+(1-\mathscr{X})\int\limits_{\Omega\times\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}g_i\left(\omega,s^{-i}|s^i\right)p_\Omega\left(\mathrm{d}\omega\right)\prod_{j\neq i}^N\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right)\\
&= \mathscr{X}\int\limits_{\Omega\times\mathbb{X}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}\int\limits_{\mathbb{S}^{-i}} g_i\left(s^{-i}|s^i\right)\prod_{j\neq i}\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right)\times g_i\left(\omega|s^i\right)p_\Omega\left(\mathrm{d}\omega\right)\\
&\quad+(1-\mathscr{X})\int\limits_{\Omega\times\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}g_i\left(\omega,s^{-i}|s^i\right)p_\Omega\left(\mathrm{d}\omega\right)\prod_{j\neq i}^N\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right)\\
&= \mathscr{X}\int\limits_{\Omega\times\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}g_i\left(s^{-i}|s^i\right)\times g_i\left(\omega|s^i\right)p_\Omega\left(\mathrm{d}\omega\right)\prod_{j\neq i}\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right)\\
&\quad+(1-\mathscr{X})\int\limits_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}}\mathrm{Prop}\left(y^i|x,s\right)\prod_{j\neq i}^N\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right)\\
&= \mathscr{X}\int\limits_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}\Omega}\left[\int\limits \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}}g_i\left(s^{-i}|s^i\right)\times g_i\left(\omega|s^i\right)p_\Omega\left(\mathrm{d}\omega\right)\right]\prod_{j\neq i}\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right)\\
&\quad+(1-\mathscr{X})\int\limits_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}}\mathrm{Prop}\left(y^i|x,s\right)\prod_{j\neq i}^N\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right),
\end{aligned}
$$

where $g_i\left(\omega, x^{-i}|s^i\right)$ is the Radon-Nikodym derivative and

$$\text{Prop}\left(y^i|x,s\right) = \int_{\Omega} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}} g_i\left(\omega, s^{-i}|s^i\right) p_\Omega\left(\mathrm{d}\omega\right).$$

The application of Theorem 4.1 requires the following steps. First, we have to define a game $\tilde{\mathscr{G}}_{OS}$, which reflects the structure of game $\mathscr{G}_{OS}$ but features players with dual-selves. As laid out in Section 4.4.1, this requires us to define (i) the space $\mathbb{X}^{\hat{i}}$, that specifies the action sets of the belief selves (ii) how these actions relate to the consequences that the action selves anticipate $Q^i_{x^{\hat{i}}}$, and (iii) the objectives of the belief selves $\pi^{\hat{i}}(\cdot)$. If all three of these objects satisfy the requirements in Theorem 4.1, there exists a Dual-Selves equilibrium in the game $\tilde{\mathscr{G}}_{OS}$. Finally, we have to argue that given $\mathbb{X}^{\hat{i}}$, $Q^i_{x^{\hat{i}}}$ and $\pi^{\hat{i}}(\cdot)$ the Dual-Selves equilibrium in game $\tilde{\mathscr{G}}_{OS}$ is indeed a Cursed Equilibrium in game $\mathscr{G}_{OS}$.

Define $\mathbb{X}^{\hat{i}}$ to be equal to Σ^{-i}, i.e., the set of all combinations of distributional strategies of action selves other than i. Moreover, define $Q^i_{x^{\hat{i}}}$ such that

$$\begin{aligned}
&Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right) \\
&= \mathscr{X} \int_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \left[\int_{\Omega} \mathbb{1}_{\{f^i(x^i,x^{-i},\omega)=y^i\}} g_i\left(s^{-i}|s^i\right) \times g_i\left(\omega|s^i\right) p_\Omega\left(\mathrm{d}\omega\right) \right] x^{\hat{i}}\left(\mathrm{d}x^{-i}, \mathrm{d}s^{-i}\right) \\
&\quad + (1-\mathscr{X}) \int_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \text{Prop}\left(y^i|x,s\right) x^{\hat{i}}\left(\mathrm{d}x^{-i}, \mathrm{d}s^{-i}\right),
\end{aligned}$$

and $\pi^{\hat{i}}$ such that

$$\pi^{\hat{i}}(\sigma, x^{\hat{i}}) = -\sum_{y^i\in\mathbb{Y}^i} (Q^i_{\bar{\sigma}}(y^i|x^i,s^i) - Q^i_{x^{\hat{i}}}(y^i|x^i,s^i))^2.$$

By the same arguments as in *Step 1a* from the proof of Theorem 4.1, the belief selves strategy sets Σ^{-i} are compact metric spaces in the weak topology. Furthermore, the set of strategies is clearly convex.

We seek to show that for any sequence with $\left(\sigma_n, x^{\hat{i}}_n\right) \Rightarrow \left(\sigma, x^{\hat{i}}\right)$ it also holds that $\pi^{\hat{i}}\left(\sigma_n, x^{\hat{i}}_n\right) \to \pi^{\hat{i}}\left(\sigma, x^{\hat{i}}\right)$.

$$\lim_{n\to\infty} \pi^{\hat{i}}\left(\sigma_n, x^{\hat{i}}_n\right) = \lim_{n\to\infty} -\sum_{y^i\in\mathbb{Y}^i} (Q^i_{\bar{\sigma}_n}(y^i|x^i,s^i) - Q^i_{x^{\hat{i}}_n}(y^i|x^i,s^i))^2$$

$$= -\sum_{y^i \in \mathbb{Y}^i} (\lim_{n\to\infty} Q^i_{\tilde{\sigma}_n}(y^i|x^i,s^i) - \lim_{n\to\infty} Q^i_{x^{\hat{i}}_n}(y^i|x^i,s^i))^2$$

As $Q^i_{\tilde{\sigma}_n}(y^i|x^i,s^i)$ and $Q^i_{x^{\hat{i}}_n}(y^i|x^i,s^i)$ are integrals over continuous and bounded functions for all y^i over the set $\mathbb{X}\times\mathbb{S}$, we can apply Theorem 1.1 in Feinberg, Kasyanov, and Zadoianchuk (2014). Consequently, $\pi^{\hat{i}}\left(\sigma, x^{\hat{i}}\right)$ is continuous. Moreover, the expected utility of action self i is given by

$$\int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{y^i\in\mathbb{Y}^i} \pi^i\left(x^i,y^i\right) Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right) \mathrm{d}\sigma^i\left(x^i,s^i\right),$$

which is continuous in the weak topology by Feinberg, Kasyanov, and Zadoianchuk (2014). We conclude that by Theorem 4.1 a Dual-Selves equilibrium exist.

In any Dual-Selves equilibrium of game $\tilde{\mathcal{G}}_{OS}$ beliefs are chosen to maximize $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu(\mathrm{d}x^{\hat{i}})$. The proposed $\pi^{\hat{i}}$, however, is non-positive and the action of the belief self that achieves $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu(\mathrm{d}x^{\hat{i}}) = 0$ is in the set $\mathbb{X}^{\hat{i}}$ for all i. In fact, in equilibrium the beliefs satisfy $Q^i_{x^{\hat{i}}}(y^i|x_i,s_i) = Q^i_{\tilde{\sigma}}(y^i|x_i,s_i)$ for all y^i, s^i, x^i and all player $i \in \{1,\cdots,N\}$. Hence, the derived strategy profile constitutes a Cursed Equilibrium in game $\mathcal{G}_{OS}$. □

4.B.2 Existence of Personal Nash Equilibrium

Next, we turn to the concept of Personal Nash Equilibrium (PNE). In analogy to Kőszegi and Rabin (2007) and Dato et al. (2017), we define the utility of a loss averse player as follows:

$$\begin{aligned}
&U^i\left(\sigma^i, \hat{\sigma}^i, \sigma^{-i}\right) \\
&= \int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{y^i\in\mathbb{Y}^i} \pi^i\left(x^i,y^i\right) Q^i_{\sigma}\left(y^i|s^i,x^i\right) \mathrm{d}\sigma^i\left(x^i,s^i\right) + \eta \int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{y^i\in\mathbb{Y}^i} Q^i_{\sigma}\left(y^i|s^i,x^i\right) \\
&\quad \cdot \left[\int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{\tilde{y}^i\in\mathbb{Y}^i} Q^i_{\tilde{\sigma}}\left(\tilde{y}^i|\tilde{s}^i,\tilde{x}^i\right) \mu\left[\pi^i\left(x^i,y^i\right) - \pi^i\left(\tilde{x}^i,\tilde{y}^i\right)\right] \tilde{\sigma}^i\left(\mathrm{d}\tilde{x}^i,\mathrm{d}\tilde{s}^i\right)\right] \sigma^i\left(\mathrm{d}x^i,\mathrm{d}s^i\right) \\
&= \int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{y^i\in\mathbb{Y}^i} \Bigg(\pi^i\left(x^i,y^i\right) + \eta\Bigg(\int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{\tilde{y}^i\in\mathbb{Y}^i} Q^i_{\tilde{\sigma}}\left(\tilde{y}^i|\tilde{s}^i,\tilde{x}^i\right) \\
&\quad \cdot \mu\left[\pi^i\left(x^i,y^i\right) - \pi^i\left(\tilde{x}^i,\tilde{y}^i\right)\right] \tilde{\sigma}^i\left(\mathrm{d}\tilde{x}^i,\mathrm{d}\tilde{s}^i\right)\Bigg)\Bigg) \cdot Q^i_{\sigma}\left(y^i|s^i,x^i\right) \mathrm{d}\sigma^i\left(x^i,s^i\right),
\end{aligned}$$

where $\hat{\sigma}^{-i}$ denotes player i's belief about the own strategy, and $\mu[\cdot]:\mathbb{R}\to\mathbb{R}$ is a continuous gain loss function. According to Dato et al. (2017) a PNE of the game is then defined as follows:

Definition 4.4. *A strategy profile σ is a PNE of game $\mathcal{G}_{OS}$ if, for all players $i\in\{1,\cdots,N\}$, we have $U^i(\sigma^i,\sigma^i)\geq U^i(\tilde{\sigma}^i,\sigma^i)$ for all $\tilde{\sigma}^i\in\Sigma^i$.*

This utility function directly depends on the players' beliefs about their own strategy in equilibrium, because their reference point is shaped by their expectations about their own future choices. Consequently, the standard assumptions of Nash-equilibrium do not apply in game $\mathcal{G}_{OS}$. However, in Dual-Selves games, we allow the belief selves' actions to influence the utility of the corresponding action self directly. We can, therefore, apply Theorem 4.1 to extend the existence result in Dato et al. (2017) to non-finite games.

Corollary 4.3. *Assume that Assumption 4.1 holds and $N=2$. Then there exists a Personal Nash Equilibrium in game $\mathcal{G}_{OS}$.*

Proof. The application of Theorem 4.1 requires the following steps. First, we have to define a game $\tilde{\mathcal{G}}_{OS}$, which reflects the structure of game $\mathcal{G}_{OS}$ but features players with dual-selves. As laid out in Section 4.4.1, this requires us to define (i) the space $\mathbb{X}^{\hat{i}}$, that specifies the action sets of the belief selves (ii) how these actions relate to the consequences that the action selves anticipate $Q^i_{x^{\hat{i}}}$, and (iii) the objectives of the belief selves $\pi^{\hat{i}}(\cdot)$. If all three of these objects satisfy the requirements in Theorem 4.1, there exists a Dual-Selves equilibrium in the game $\tilde{\mathcal{G}}_{OS}$. Finally, we have to argue that given $\mathbb{X}^{\hat{i}}$, $Q^i_{x^{\hat{i}}}$ and $\pi^{\hat{i}}(\cdot)$ the Dual-Selves equilibrium in game $\tilde{\mathcal{G}}_{OS}$ is indeed a PNE in game $\mathcal{G}_{OS}$.

Define $\mathbb{X}^{\hat{i}}$ to be equal to Σ, where Σ is the set of distributional strategies. Moreover, define $Q^i_{x^{\hat{i}}}$ such that

$$Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)=\int_{\mathbb{X}\times\mathbb{S}}\text{Prop}\left(y^i|x,s\right)x^{\hat{i}}\left(\mathrm{d}x^{-i},\mathrm{d}s^{-i}\right)$$

and $\pi^{\hat{i}}$ as follows:

$$\pi^{\hat{i}}(\sigma,x^{\hat{i}})=-\sum_{y\in\mathbb{Y}^i}(Q^i_{x^{\hat{i}}}(y|x_i,s_i)-Q^i_{\sigma}(y|x_i,s_i)^i)^2$$

The utility of the action self given the belief $x^{\hat{i}}$ is then given by:

$$\begin{aligned}
&U^i\left(\sigma^i, x^{\hat{i}}\right) \\
&= \int_{\mathbb{X}^i \times \mathbb{S}^i} \sum_{y^i \in \mathbb{Y}^i} \pi^i\left(x^i, y^i\right) Q^i_{x^{\hat{i}}}\left(y^i | s^i, x^i\right) \mathrm{d}\sigma^i\left(x^i, s^i\right) + \eta \int_{\mathbb{X}^i \times \mathbb{S}^i} \sum_{y^i \in \mathbb{Y}^i} Q^i_{x^{\hat{i}}}\left(y^i | s^i, x^i\right) \\
&\quad \cdot \left[\int_{\mathbb{X}^i \times \mathbb{S}^i} \sum_{\tilde{y}^i \in \mathbb{Y}^i} Q^i_{x^{\hat{i}}}\left(\tilde{y}^i | \tilde{s}^i, \tilde{x}^i\right) \mu\left[\pi^i\left(x^i, y^i\right) - \pi^i\left(\tilde{x}^i, \tilde{y}^i\right) \right] x^{\hat{i}}\left(\mathrm{d}\tilde{x}^i, \mathrm{d}\tilde{s}^i\right) \right] \sigma^i\left(\mathrm{d}x^i, \mathrm{d}s^i\right)
\end{aligned}$$

Note that we do not need and, therefore, have dropped $x^{\hat{i}}$ as an argument of the payoff function of action self i. By the same arguments as in *Step 1a* from the proof of Theorem 4.1, the belief selves strategy sets Σ are compact metric spaces in the weak topology. Furthermore, the set of strategies is clearly convex.

The continuity of $Q^i_{x^{\hat{i}}}$ on $\mathbb{X}^{\hat{i}} \times \mathbb{S}^i \times \mathbb{X}^i$ and the continuity of $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}}) \mu(\mathrm{d}x^{\hat{i}})$ on $\mathscr{M}(\mathbb{X}^{\hat{i}}) \times \Sigma$ follow from the same arguments as in Section 4.B.1.

Next, we turn to the continuity of $U^i(\sigma^i, x^{\hat{i}})$ in the weak topology. For any sequence $\lim_{n\to\infty}(\sigma^i_n, x^{\hat{i}}_n) \Rightarrow (\sigma^i, x^{\hat{i}})$, we get:

$$\begin{aligned}
&\lim_{n\to\infty} U^i\left(\sigma^i_n, x^{\hat{i}}_n\right) \\
&= \lim_{n\to\infty} \int_{\mathbb{X}^i \times \mathbb{S}^i} \sum_{y^i \in \mathbb{Y}^i} \pi^i\left(x^i, y^i\right) Q^i_{x^{\hat{i}}_n}\left(y^i | s^i, x^i\right) \mathrm{d}\sigma^i_n\left(x^i, s^i\right) + \lim_{n\to\infty} \eta \int_{\mathbb{X}^i \times \mathbb{S}^i} \sum_{y^i \in \mathbb{Y}^i} Q^i_{x^{\hat{i}}_n}\left(y^i | s^i, x^i\right) \\
&\quad \cdot \left[\int_{\mathbb{X}^i \times \mathbb{S}^i} \sum_{\tilde{y}^i \in \mathbb{Y}^i} Q^i_{x^{\hat{i}}_n}\left(\tilde{y}^i | \tilde{s}^i, \tilde{x}^i\right) \mu\left[\pi^i\left(x^i, y^i\right) - \pi^i\left(\tilde{x}^i, \tilde{y}^i\right) \right] x^{\hat{i}}_n\left(\mathrm{d}\tilde{x}^i, \mathrm{d}\tilde{s}^i\right) \right] \sigma^i_n\left(\mathrm{d}x^i, \mathrm{d}s^i\right)
\end{aligned}$$

Making use of the insights from above and Assumption 4.1, we can rewrite Q^i_σ and $Q^i_{x^{\hat{i}}}$ such that:

$$Q^i_\sigma\left(y^i | s^i, x^i\right) = \int_{\mathbb{X}^{-i} \times \mathbb{S}^{-i}} \mathrm{Prop}\left(y^i | x, s\right) \prod_{j \neq i}^{N} \sigma^j\left(\mathrm{d}x^j, \mathrm{d}s^j\right)$$

and

$$Q^i_{x^{\hat{i}}}\left(y^i | s^i, x^i\right) = \int_{\mathbb{X}^{-i} \times \mathbb{S}^{-i}} \mathrm{Prop}\left(y^i | x, s\right) \prod_{j \neq i}^{N} x^{\hat{i}}_j\left(\mathrm{d}x^j, \mathrm{d}s^j\right),$$

where $x^{\hat{i}}_j$ denotes the belief self $\hat{i}$'s belief about the strategy of action self j. The continuity and boundedness of $\mathrm{Prop}(y_i | x, s)$ implies that both probability distributions are

continuous and bounded. Applying Feinberg, Kasyanov, and Zadoianchuk (2014), we get:

$$\begin{aligned}
&\lim_{n\to\infty} U^i\left(\sigma_n^i, x_n^{\hat{i}}\right) \\
&= \int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{y^i\in\mathbb{Y}^i} \pi^i\left(x^i,y^i\right) Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right) \mathrm{d}\sigma^i\left(x^i,s^i\right) + \lim_{n\to\infty} \eta \int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{y^i\in\mathbb{Y}^i} Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right) \\
&\quad \cdot \left[\int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{\tilde{y}^i\in\mathbb{Y}^i} Q^i_{x^{\hat{i}}}\left(\tilde{y}^i|\tilde{s}^i,\tilde{x}^i\right) \mu\left[\pi^i\left(x^i,y^i\right) - \pi^i\left(\tilde{x}^i,\tilde{y}^i\right)\right] x_n^{\hat{i}}\left(\mathrm{d}\tilde{x}^i,\mathrm{d}\tilde{s}^i\right)\right] \sigma_n^i\left(\mathrm{d}x^i,\mathrm{d}s^i\right).
\end{aligned}$$

To exchange limits, we have to show that the function in the integral is bounded and continuous on $\mathbb{X}^i \times \mathbb{S}^i$. For every sequence $(x_n^i, s_n^i) \to (x^i, s^i)$, we have

$$\begin{aligned}
\lim_{n\to\infty} \sum_{y^i\in\mathbb{Y}^i} Q^i_{x^{\hat{i}}}\left(y^i|s_n^i,x_n^i\right) \Bigg[\int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{\tilde{y}^i\in\mathbb{Y}^i} Q^i_{x^{\hat{i}}}\left(\tilde{y}^i|\tilde{s}^i,\tilde{x}^i\right) \mu\Big[\pi^i\left(x_n^i,y^i\right) \\
-\pi^i\left(\tilde{x}^i,\tilde{y}^i\right)\Big] x^{\hat{i}}\left(\mathrm{d}\tilde{x}^i,\mathrm{d}\tilde{s}^i\right)\Bigg],
\end{aligned}$$

which results in

$$\sum_{y^i\in\mathbb{Y}^i} Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right) \left[\int_{\mathbb{X}^i\times\mathbb{S}^i} \sum_{\tilde{y}^i\in\mathbb{Y}^i} Q^i_{x^{\hat{i}}}\left(\tilde{y}^i|\tilde{s}^i,\tilde{x}^i\right) \mu\left[\pi^i\left(x^i,y^i\right) - \pi^i\left(\tilde{x}^i,\tilde{y}^i\right)\right] x^{\hat{i}}\left(\mathrm{d}\tilde{x}^i,\mathrm{d}\tilde{s}^i\right)\right],$$

since the sum and multiplication of continuous and bounded functions is continuous and bounded. Consequently we get

$$\lim_{n\to\infty} U^i\left(\sigma_n^i, x_n^{\hat{i}}\right) = U^i\left(\sigma^i, x^{\hat{i}}\right).$$

Therefore, a Dual-Selves equilibrium in game $\tilde{\mathcal{G}}_{OS}$ exists.

In any Dual-Selves equilibrium in game $\tilde{\mathcal{G}}_{OS}$ beliefs are chosen to maximize $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu(\mathrm{d}x^{\hat{i}})$. The proposed $\pi^{\hat{i}}$, however, is non positive and the belief that achieves $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu(\mathrm{d}x^{\hat{i}}) = 0$ is in the set $\mathbb{X}^{\hat{i}}$ for all i. In fact, in equilibrium the beliefs satisfy $Q^i_{x^{\hat{i}}}(y|x_i, s_i) = Q^i_{\sigma}(y|x_i, s_i)$ for all s_i, x_i and all action selves $i \in \{1, \cdot, N\}$. Hence, the derived strategy profile constitutes a PNE in game $\mathcal{G}_{OS}$. □

4.B.3 Existence of Psychological Equilibrium

Next, consider the concept of psychological games. Take game $\mathscr{G}_{OS}$ in which each player only has one self. In analogy to Geanakoplos, Pearce, and Stacchetti (1989), define the beliefs of these players as follows. A first order belief is a probability measure over the product of the strategy sets of all other players. Hence, the set of first order beliefs of player i is given by $B_1^i = \mathscr{M}(\Sigma)$. The sets of higher order beliefs are then inductively defined as:

$$B_k^i = \mathscr{M}(\Sigma_{-i} \times B_1^{-i} \times \cdot \times B_k^{-i})$$
$$B_k^{-i} = \times_{j \neq i} B_k^j \qquad B_k = \times_i B_k^i$$
$$B^i = \times_{k \in \{1,\cdots,K\}} B_k^i \qquad B = \times_i B^i$$

For simplicity, we assume that the payoff relevant beliefs are of finite order K. Player i's payoff if he holds belief $b^i \in B^i$, he plays x^i and outcome y^i is realized is then $\pi^i(x^i, y^i, b^i) : \mathbb{X}^i \times \mathbb{Y}^i \times B^i \to \mathbb{R}$ and expected utility is given by

$$U^i\left(b^i, \sigma\right) = \int_{\mathbb{X}^i \times \mathbb{S}^i} Q_\sigma^i(y^i | s^i, x^i)\, \pi^i(x^i, y^i, b^i)\, \sigma^i(\mathrm{d}x^i, \mathrm{d}s^i).$$

We make the following assumption on the game $\mathscr{G}_{OS}$ to transfer the notion of a Psychological Equilibrium to our setting.

Assumption 4.4. *$\pi^i(x^i, y^i, b^i)$ is continuous on $\mathbb{X}^i \times \mathbb{Y}^i \times B^i$.*

Geanakoplos, Pearce, and Stacchetti (1989) require that beliefs about other players' actions are correct in equilibrium. If σ is an equilibrium profile of strategies, therefore, all players expect their opponents to play σ^{-i}, and that each opponent j expects that all other players play σ^{-j} and so on. Denote the corresponding belief system by $\beta(\sigma) = \left(\beta_1(\sigma), \cdots, \beta_n(\sigma)\right) \in B$. According to Geanakoplos, Pearce, and Stacchetti (1989) a Psychological Equilibrium is then defined as follows:

Definition 4.5. *A profile $(b, \sigma) \in B \times \Sigma$ is a Psychological Equilibrium of $\mathscr{G}_{OS}$ iff:*

(i) $b = \beta(\sigma)$.

(ii) $\forall\ i \in \{1, \cdots, N\}$, *we have* $U^i(b^i, (\sigma^i, \sigma^{-i})) \geq U^i(b^i, (\tilde{\sigma}^i, \sigma^{-i}))$ *for all* $\tilde{\sigma}^i \in \Sigma^i$.

The following corollary extends the result in Geanakoplos, Pearce, and Stacchetti (1989) to non-finite games.

Corollary 4.4. *Suppose that assumption 4.1, 4.2 & 4.4 hold and $N = 2$, then there exist a Psychological Equilibrium in $\mathscr{G}_{OS}$.*

Proof. The application of Theorem 4.1 requires the following steps. First, we have to define a game $\tilde{\mathscr{G}}_{OS}$, which reflects the structure of game $\mathscr{G}_{OS}$ but features players with dual-selves. As laid out in Section 4.4.1, this requires us to define (i) the space $\mathbb{X}^{\hat{i}}$, that specifies the action sets of the belief selves (ii) how these actions relate to the consequences that the action selves anticipate $Q^i_{x^{\hat{i}}}$, and (iii) the objectives of the belief selves $\pi^{\hat{i}}(\cdot)$. If all three of these objects satisfy the requirements in Theorem 4.1, then there exists a Dual-Selves equilibrium in the game $\tilde{\mathscr{G}}_{OS}$. Finally, we have to argue that given $\mathbb{X}^{\hat{i}}$, $Q^i_{x^{\hat{i}}}$ and $\pi^{\hat{i}}(\cdot)$ the Dual-Selves equilibrium in game $\tilde{\mathscr{G}}_{OS}$ is indeed a Psychological Equilibrium in game $\mathscr{G}_{OS}$. Define $\mathbb{X}^{\hat{i}}$ to be equal to B^i and $\pi^{\hat{i}}$ as follows:

$$\pi^{\hat{i}}(\sigma, x^{\hat{i}}) = -d(x^{\hat{i}}, \beta(\sigma)),$$

where $d(\cdot,\cdot)$ is a bounded metric on B, which exists and is continuous because B^i is a metric space.[12]

The action spaces are assumed to be compact, and, hence, they and their Cartesian-product are also complete (see Theorem 8.16, p.96, in De la Fuente (2000)). By Theorem 3.2 in Parthasarathy (1967), Σ is then a tight set of probability measures. Also, since it is a set of measures over separable and complete metric spaces, it is complete itself by Theorem 6.5 in Parthasarathy (1967) and, hence, closed in the weak topology. With the same argument, the sets B^i_k are closed because they are the Cartesian product of sets of measures over complete and compact metric spaces. By Prohorov's Theorem, it follows that the strategy sets Σ, and the belief sets B^i are compact metric spaces in the weak topology. Furthermore, the set of strategies is clearly convex.

12. See Theorem 20.1 in Munkres (2000).

Since the metric is continuous and bounded on $B^i \times B^i$ also $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu(\mathrm{d}x^{\hat{i}})$ is continuous on $\mathscr{M}(\mathbb{X}^{\hat{i}}) \times \Sigma$. Moreover, the utility functions are continuous since $Q^i_\sigma(y^i|s^i,x^i)\pi^i(x^i,y^i,b^i)$ is continuous and bounded (for a proof see the proofs of Corollaries 4.3 and 4.5. Therefore, a Dual-Selves equilibrium in game $\tilde{\mathscr{G}}_{OS}$ exists.

In any Dual-Selves equilibrium in game $\tilde{\mathscr{G}}_{OS}$ beliefs are chosen to maximize $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu(\mathrm{d}x^{\hat{i}})$. The proposed $\pi^{\hat{i}}$, however, is non-positive and the belief that achieves $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu(\mathrm{d}x^{\hat{i}}) = 0$ is in the set $\mathbb{X}^{\hat{i}}$ for all i. In fact, in equilibrium the belief needs to satisfy $b^i = \beta^i(\sigma)$ for all action selves $i \in \{1,\dots,N\}$. Hence, the derived strategy profile has to constitute a Psychological Equilibrium. □

4.B.4 Existence of Berk-Nash Equilibrium

As another example, our result extends earlier results to non-finite games considering the concept of a Berk-Nash Equilibrium by Esponda and Pouzo (2016). To transfer the notion of a Berk-Nash Equilibrium to our setting, consider the game $\mathscr{G}_{OS}$ in which every player has only one self. As the optimality criterion that determines how beliefs about primitives and opponents' strategies are chosen, Esponda and Pouzo (2016) make use of the weighted Kullback-Leibler divergence (wKLD). Transferred to our notation, the weighted Kullback-Leibler divergence for action self i is given by

$$K^i\left(\sigma, x^{\hat{i}}\right) = \int_{s^i \in \mathbb{S}^i} \int_{x^i \in \mathbb{X}^i} E_{Q^i_\sigma(\cdot|s^i,x^i)} \left[\ln \frac{Q^i_\sigma\left(y^i|s^i,x^i\right)}{Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)} \right] \sigma^i\left(\mathrm{d}x^i|s^i\right) \cdot p_{S^i}\left(\mathrm{d}s^i\right), \tag{4.B.1}$$

where $x^{\hat{i}} \in \mathbb{X}^{\hat{i}}$ reflects the model that player i has formed about the game. Essentially, this model summarizes what player i beliefs about the structure of the game and the strategies that his opponents will play (see Esponda and Pouzo, 2016). We define the set of parameter values $\mathbb{X}^{\hat{i}}(\sigma) \subset \mathbb{X}^{\hat{i}}$ to be the set of all beliefs such that the wKLD is minimized:

$$\mathbb{X}^{\hat{i}}(\sigma) \equiv arg\min_{x^{\hat{i}} \in \mathbb{X}^{\hat{i}}} K^i\left(\sigma, x^{\hat{i}}\right).$$

The interpretation is that $\mathbb{X}^{\hat{i}}(\sigma)$ is the set of beliefs inducing probability distributions over outcomes that best match the objective distribution over outcomes. A Berk-Nash equilibrium requires each player to choose a strategy that is optimal given her beliefs. A strategy σ^i for player i is optimal given $\mu^i \in \mathscr{M}\left(\mathbb{X}^{\hat{i}}\right)$ if $\sigma^i\left(x^i, s^i\right) > 0$ implies that

$$x^i \in arg\max_{\tilde{x}^i \in \mathbb{X}^i} E_{\bar{Q}^i_{\mu^i}(\cdot|s^i,\tilde{x}^i)}\left[\pi^i\left(\tilde{x}^i, y^i\right)\right], \tag{4.B.2}$$

where $\bar{Q}^i_{\mu^i}\left(\cdot|s^i,x^i\right) = \int_{X^{\hat{i}}} Q^i_{x^{\hat{i}}}\left(\cdot|s^i,x^i\right)\mu^i\left(\mathrm{d}x^{\hat{i}}\right)$ is the distribution over consequences of player i, conditional on $\left(s^i,x^i\right) \in \mathbb{S}^i \times \mathbb{X}^i$, induced by μ^i. According to Esponda and Pouzo (2016), a Berk-Nash-Equilibrium is defined as follows.

Definition 4.6. *A strategy profile σ is a Berk-Nash equilibrium of game $\mathscr{G}_{OS}$ if, for all players $i \in I$, there exists $\mu^i \in \mathscr{M}\left(\mathbb{X}^{\hat{i}}\right)$ such that (i) σ^i is optimal given μ^i, and (ii) $\mu^i \in \mathscr{M}\left(\mathbb{X}^{\hat{i}}(\sigma)\right)$, that is, if $\hat{x}^i$ is in the support of μ^i, then $\hat{x}^i \in arg\min_{x^{\hat{i}} \in \mathbb{X}^{\hat{i}}} K^i\left(\sigma, x^{\hat{i}}\right)$*

Similar to Esponda and Pouzo (2016), we maintain the following assumption about feasible beliefs.

Assumption 4.5. *For all $i \in I$ and for all $\sigma^{-i} \in \Sigma^{-i}$, $y^i \in \mathbb{Y}^i$ and $(x^{\hat{i}}_n, s^i_n, x^i_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty}(x^{\hat{i}}_n, s^i_n, x^i_n) = (x^{\hat{i}}, s^i, x^i)$ there exists $\epsilon > 0$ such that if $Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right) < \epsilon$, there exist $n' \in \mathbb{N}$ and $m \in \mathbb{N}$ with $m < \infty$ such that $Q_\sigma\left(y^i|s^i_{n''},x^i_{n''}\right)^m \leq Q^i_{x^{\hat{i}}_{n''}}\left(y^i|s^i_{n''},x^i_{n''}\right)$ for all $n'' > n'$.*

This assumption restricts misspecified beliefs that attach zero probability to some outcomes given other players' strategies in two ways. First, it requires that no misspecification renders a consequence impossible if it, in fact, occurs with positive probability. Second, we also impose a similar restriction on consequences in an ϵ-environment of subjectively impossible consequences: When approaching a subjective model that renders a consequence impossible, the subjective probability must not converge zero at a much faster rate than the objective probability. While the latter is trivially fulfilled in any discrete game, the first part reduces the set of covered misspecifications compared to Esponda and Pouzo (2016). For instance, it rules out the belief that other action selves only follow pure strategies in some discrete games.

Nevertheless, a broad class of models fulfills Assumption 4.5. For example, all combinations of games and misspecified beliefs in which players attach positive subjective probabilities to every consequence fulfill this assumption.

Corollary 4.5. *Suppose that Assumptions 4.1, 4.2 & 4.5 hold, then a Berk-Nash Equilibrium exists.*

Proof. The application of Theorem 4.1 requires to define a game $\tilde{\mathcal{G}}_{OS}$, which reflects the structure of game $\mathcal{G}_{OS}$ but features players with dual-selves. As laid out in Section 4.4.1, this requires us to define $\pi^{\hat{i}}$ additionally to the set of actions of the belief selves which is given by the set of potential models $\mathbb{X}^{\hat{i}}$. Take $\pi^{\hat{i}}\left(\sigma, x^{\hat{i}}\right) = -K^i\left(\sigma, x^{\hat{i}}\right)$. Then from Definition 4.6 we can infer that a Dual-Selves equilibrium in game $\tilde{\mathcal{G}}_{OS}$ constructs a Berk-Nash Equilibrium in $\mathcal{G}_{OS}$. Therefore, it suffices to show that we can apply Theorem 4.1. We can do so if $K^i\left(\sigma, x^{\hat{i}}\right)$ is continuous.

Therefore, we seek to show that for any sequence with $\left(\sigma_n, \mu_n^{\hat{i}}\right) \to \left(\sigma, \mu^{\hat{i}}\right)$ it also holds that $\int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma_n, x^{\hat{i}})\mu_n^{\hat{i}}(\mathrm{d}x^{\hat{i}}) \to \int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma, x^{\hat{i}})\mu^{\hat{i}}(\mathrm{d}x^{\hat{i}})$. We get

$$\lim_{n\to\infty} \int_{\mathbb{X}^{\hat{i}}} \pi^{\hat{i}}(\sigma_n, x^{\hat{i}})\mu_n^{\hat{i}}(\mathrm{d}x^{\hat{i}}) \tag{4.B.3}$$

$$= \lim_{n\to\infty} -\int_{\mathbb{X}^{\hat{i}}} K^i\left(\sigma_n, x^{\hat{i}}\right)\mu_n^{\hat{i}}\left(\mathrm{d}x^{\hat{i}}\right)$$

$$= \lim_{n\to\infty} -\int_{\mathbb{X}^{\hat{i}}\times\mathbb{S}^i\times\mathbb{X}^i} E_{Q^i_{\sigma_n}(\cdot|s^i,x^i)}\left[\ln \frac{Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)}{Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)}\right] \sigma_n^i\left(\mathrm{d}x^i|s^i\right)\cdot p_{S^i}\left(\mathrm{d}s^i\right)\mu_n^{\hat{i}}\left(\mathrm{d}x^{\hat{i}}\right)$$

$$= \lim_{n\to\infty} \int_{\mathbb{X}^{\hat{i}}\times\mathbb{S}^i\times\mathbb{X}^i} -\sum_{y^i\in\mathbb{Y}^i} Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\ln Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\sigma_n^i\left(\mathrm{d}x^i,\mathrm{d}s^i\right)\mu_n^{\hat{i}}\left(\mathrm{d}x^{\hat{i}}\right) \tag{4.B.4}$$

$$+ \lim_{n\to\infty} \int_{\mathbb{X}^{\hat{i}}\times\mathbb{S}^i\times\mathbb{X}^i} \sum_{y^i\in\mathbb{Y}^i} Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\ln Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)\sigma_n^i\left(\mathrm{d}x^i,\mathrm{d}s^i\right)\mu_n^{\hat{i}}\left(\mathrm{d}x^{\hat{i}}\right). \tag{4.B.5}$$

First, consider the term in (4.B.4). To derive continuity on $\mathbb{X}^i \times \mathbb{S}^i$, we first start by arguing that the integral and the limit in (4.B.4) can be exchanged. For this purpose, we show that $\sum_{y^i\in\mathbb{Y}^i} -Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\ln Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)$ is bounded and continuous on $\mathbb{X}^i \times \mathbb{S}^i$. For any sequence (x_n^i, s_n^i) with $\lim_{n\to\infty}(x_n^i, s_n^i) = (x^i, s^i)$, we get

$$\lim_{n\to\infty} Q^i_\sigma\left(y^i|s^i_n,x^i_n\right) = \lim_{n\to\infty}\int_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \text{Prop}\left(y^i|(x^i_n,x^{-i}),(s^i_n,s^{-i})\right)\prod_{j\neq i}^{N}\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right).$$

Since $\text{Prop}\left(y^i|(x^i_n,x^{-i}),(s^i_n,s^{-i})\right)$ is bounded and continuous on $\mathbb{X}\times\mathbb{S}\ \ \forall\, y^i$ and i, we can apply Theorem 1.1 in Feinberg, Kasyanov, and Zadoianchuk (2014) and get

$$\begin{aligned}
&\lim_{n\to\infty}\int_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \text{Prop}\left(y^i|(x^i_n,x^{-i}),(s^i_n,s^{-i})\right)\prod_{j\neq i}^{N}\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right)\\
&=\int_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \text{Prop}\left(y^i|(x^i,x^{-i}),(s^i,s^{-i})\right)\prod_{j\neq i}^{N}\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right)\\
&=Q^i_\sigma\left(y^i|s^i,x^i\right),
\end{aligned}$$

which implies continuity of $Q^i_\sigma\left(y^i|s^i,x^i\right)$. It is also bounded as $Q^i_\sigma\left(y^i|s^i,x^i\right)\in[0,1]$ holds. Hence, the function $\sum_{y^i\in\mathbb{Y}^i} -Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\ln Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)$ is bounded and continuous. By Theorem 1.1 in Feinberg, Kasyanov, and Zadoianchuk (2014) and the continuity of $Q^i_\sigma\left(y^i|s^i,x^i\right)$ on $\mathbb{X}^i\times\mathbb{S}^i$, we can thus exchange the limit and the integral in (4.B.4).

Next, we have to show that $\sigma_n\Rightarrow\sigma$ implies $Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\to Q^i_\sigma\left(y^i|s^i,x^i\right)$. Since $\mathbb{X}^i$ and $\mathbb{X}^{-i}$ are compact metric spaces they are complete, by Theorem 8.16, p.96, in De la Fuente (2000), and separable. Furthermore, $\mathbb{S}^i$ and $\mathbb{S}^{-i}$ are complete separable metric spaces such that $\mathbb{X}^i\times\mathbb{S}^i$ and $\mathbb{X}^{-i}\times\mathbb{S}^{-i}$ are complete and separable. Hence, we can apply Theorem 2.8 from Billingsley (1999) and deduce that $\sigma_n\Rightarrow\sigma$ implies $\sigma^{-i}_n\Rightarrow\sigma^{-i}$. Using this insight, we can apply Theorem 1.1 in Feinberg, Kasyanov, and Zadoianchuk (2014) because $\text{Prop}\left(y^i|x,s\right)$ is a bounded and continuous function on the space $\mathbb{X}^{-i}\times\mathbb{S}^{-i}\ \ \forall\, y^i$ and i:

$$\lim_{n\to\infty}\int_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \text{Prop}\left(y^i|x,s\right)\prod_{j\neq i}^{N}\sigma^j_n\left(\mathrm{d}x^j,\mathrm{d}s^j\right) = \int_{\mathbb{X}^{-i}\times\mathbb{S}^{-i}} \text{Prop}\left(y^i|x,s\right)\prod_{j\neq i}^{N}\sigma^j\left(\mathrm{d}x^j,\mathrm{d}s^j\right),$$

which is equivalent to $\lim_{n\to\infty} Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)=Q^i_\sigma\left(y^i|s^i,x^i\right)$. Hence,

$$\begin{aligned}
&\lim_{n\to\infty}\int_{\mathbb{X}^{\hat{i}}\times\mathbb{S}^i\times\mathbb{X}^i} -\sum_{y^i\in\mathbb{Y}^i} Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\ln Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\sigma^i_n\left(\mathrm{d}x^i,\mathrm{d}s^i\right)\mu^{\hat{i}}_n\left(\mathrm{d}x^{\hat{i}}\right)\\
&=-\int_{\mathbb{X}^{\hat{i}}\times\mathbb{S}^i\times\mathbb{X}^i}\sum_{y^i\in\mathbb{Y}^i} Q^i_\sigma\left(y^i|s^i,x^i\right)\ln Q^i_\sigma\left(y^i|s^i,x^i\right)\sigma^i\left(\mathrm{d}x^i,\mathrm{d}s^i\right)\mu^{\hat{i}}\left(\mathrm{d}x^{\hat{i}}\right).
\end{aligned}$$

Now, consider the term in (4.B.5). Similar to before, we will argue that $Q^i_\sigma\left(y^i|s^i,x^i\right)\ln Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)\equiv\pi^{\hat{i}}(x^{\hat{i}},s^i,x^i)$ is bounded as well as continuous on $\mathbb{X}^{\hat{i}}\times\mathbb{X}^i\times\mathbb{S}^i$. We start by proving continuity of $\pi^{\hat{i}}(\cdot)$. Consider any $(x^{\hat{i}},s^i,x^i)$ with $Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)>0$. Since $Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)$ is continuous and strictly larger than zero also $lnQ^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)$ is continuous at $(x^{\hat{i}},s^i,x^i)$. As $Q^i_\sigma\left(y^i|s^i,x^i\right)$ is continuous, $\pi^{\hat{i}}(\cdot)$ is continuous at $(x^{\hat{i}},s^i,x^i)$. Next, consider any $(x^{\hat{i}},s^i,x^i)$ with $Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)=0$. As a preliminary step we show that also $\pi^{\hat{i}}\left(y^i|s^i,x^i\right)=0$ holds. Take any sequence $(x^{\hat{i}}_n,s^i_n,x^i_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty}(x^{\hat{i}}_n,s^i_n,x^i_n)=(x^{\hat{i}},s^i,x^i)$. Since for all $n''>n'$ we have $0\leq Q^i_\sigma\left(y^i|s^i_{n''},x^i_{n''}\right)^m\leq Q^i_{x^{\hat{i}}_{n''}}\left(y^i|s^i_{n''},x^i_{n''}\right)$ such that $0\leq\lim_{n\to\infty}Q^i_\sigma\left(y^i|s^i_n,x^i_n\right)^m\leq\lim_{n\to\infty}Q^i_{x^{\hat{i}}_n}\left(y^i|s^i_n,x^i_n\right)$, and, therefore, $0\leq\lim_{n\to\infty}Q^i_\sigma\left(y^i|s^i_n,x^i_n\right)^m\leq 0$. As in Esponda and Pouzo (2016) we define $ln\,(0)=-\infty$ and $0\cdot\infty=0$, which implies $\pi^{\hat{i}}(x^{\hat{i}},s^i,x^i)=0$.

Next, we show that $\pi^{\hat{i}}(x^{\hat{i}},s^i,x^i)$ is bounded. Obviously, $\pi^{\hat{i}}(x^{\hat{i}},s^i,x^i)$ is bounded above by zero. Suppose first that $Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)\geq\epsilon$. Note that $Q^i_\sigma\left(y^i|s^i,x^i\right)\leq 1$ and $\ln Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)\leq 0$. Hence,

$$0\;\geq\;\pi^{\hat{i}}\left(x^{\hat{i}},s^i,x^i\right)\;=\;Q^i_\sigma\left(y^i|s^i,x^i\right)\cdot\ln Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)\;\geq\;\ln Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)\;\geq\;1-\frac{1}{\epsilon}.$$

Consider $Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)<\epsilon$, now. As $\pi^{\hat{i}}(x^{\hat{i}},s^i,x^i)$ is continuous at any $(x^{\hat{i}},s^i,x^i)$,

$$\begin{aligned}0\;\geq\;\pi^{\hat{i}}\left(x^{\hat{i}},s^i,x^i\right)&=\lim_{n\to\infty}\pi^{\hat{i}}(x^{\hat{i}}_n,s^i_n,x^i_n)\\&\geq\lim_{n\to\infty}m\cdot Q^i_\sigma\left(y^i|s^i_n,x^i_n\right)\cdot\ln Q^i_\sigma\left(y^i|s^i_n,x^i_n\right)\\\Leftrightarrow\;0\;\geq\;\pi^{\hat{i}}\left(x^{\hat{i}},s^i,x^i\right)&\geq m\cdot Q^i_\sigma\left(y^i|s^i,x^i\right)\cdot\ln Q^i_\sigma\left(y^i|s^i,x^i\right)\\&\geq m\left(Q^i_\sigma\left(y^i|s^i,x^i\right)-1\right)\end{aligned}$$

holds. Since $m<\infty$, $\pi^{\hat{i}}(\cdot)$ is bounded. By Theorem 1.1 in Feinberg, Kasyanov, and Zadoianchuk (2014) we can write

$$\begin{aligned}&\lim_{n\to\infty}\int_{\mathbb{X}^{\hat{i}}\times\mathbb{S}^i\times\mathbb{X}^i}-\sum_{y^i\in\mathbb{Y}^i}Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\ln Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)\sigma^i_n\left(\mathrm{d}x^i,\mathrm{d}s^i\right)\mu^{\hat{i}}_n\left(\mathrm{d}x^{\hat{i}}\right)\\&=\int_{\mathbb{X}^{\hat{i}}\times\mathbb{S}^i\times\mathbb{X}^i}-\sum_{y^i\in\mathbb{Y}^i}Q^i_\sigma\left(y^i|s^i,x^i\right)\ln Q^i_{x^{\hat{i}}}\left(y^i|s^i,x^i\right)\sigma^i\left(\mathrm{d}x^i,\mathrm{d}s^i\right)\mu^{\hat{i}}\left(\mathrm{d}x^{\hat{i}}\right),\end{aligned}$$

which also makes use of the fact that $\sigma_n\Rightarrow\sigma$ implies $Q^i_{\sigma_n}\left(y^i|s^i,x^i\right)\to Q^i_\sigma\left(y^i|s^i,x^i\right)$. Putting the limits of (4.B.4) and (4.B.5) together yields $\lim_{n\to\infty}\int_{\mathbb{X}^{\hat{i}}}\pi^{\hat{i}}(\sigma_n,x^{\hat{i}})\mu^{\hat{i}}_n(\mathrm{d}x^{\hat{i}})=\int_{\mathbb{X}^{\hat{i}}}\pi^{\hat{i}}(\sigma,x^{\hat{i}})\mu^{\hat{i}}(\mathrm{d}x^{\hat{i}})$, i.e., $\int_{\mathbb{X}^{\hat{i}}}\pi^{\hat{i}}(\sigma,x^{\hat{i}})\mu^{\hat{i}}(\mathrm{d}x^{\hat{i}})$ is continuous in the weak topology. □

References

Alonso, Ricardo, Isabelle Brocas, and Juan D. Carrillo. 2013. "Resource Allocation in the Brain." *Review of Economic Studies* 81 (2): 501–34. DOI: https://doi.org/10.1093/restud/rdt043. [122, 129]

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny. 1998. "A Model of Investor Sentiment." *Journal of Financial Economics* 49 (3): 307–43. DOI: https://doi.org/10.1016/S0304-405X(98)00027-0. [134]

Bénabou, Roland, and Jean Tirole. 2006a. "Belief in a Just World and Redistributive Politics." *Quarterly Journal of Economics* 121 (2): 699–746. DOI: https://doi.org/10.1162/qjec.2006.121.2.699. [121]

Bénabou, Roland, and Jean Tirole. 2006b. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78. DOI: https://doi.org/10.1257/aer.96.5.1652. [136]

Bénabou, Roland, and Jean Tirole. 2011. " Identity, Morals, and Taboos: Beliefs as Assets." *Quarterly Journal of Economics* 126 (2): 805–55. DOI: https://doi.org/10.1093/qje/qjr002. [121, 134]

Benjamin, Dan, AL Bodoh-Creed, and Matthew Rabin. 2019. "Base-Rate Neglect: Foundations and Implications." *Discussion Paper*, URL: https://cutt.ly/bhQKbTw. [123, 137]

Benjamin, Daniel J., Matthew Rabin, and Collin Raymond. 2016. "A Model of Nonbelief in the Law of Large Numbers." *Journal of the European Economic Association* 14 (2): 515–44. DOI: https://doi.org/10.1111/jeea.12139. [134]

Billingsley, Patrick. 1968. *Convergence of Probability Measures*. John Wiley & Sons. [139]

Billingsley, Patrick. 1999. *Convergence of Probability Measures*. John Wiley & Sons. [154]

Brocas, Isabelle, and Juan D. Carrillo. 2008. "The Brain as a Hierarchical Organization." *American Economic Review* 98 (4): 1312–46. DOI: https://doi.org/10.1257/aer.98.4.1312. [121, 122, 129]

Brunnermeier, Markus K., and Jonathan A. Parker. 2005. "Optimal Expectations." *American Economic Review* 95 (4): 1092–118. DOI: https://doi.org/10.1257/0002828054825493. [121]

Caplin, Andrew, and John Leahy. 2001. "Psychological Expected Utility Theory and Anticipatory Feelings." *Quarterly Journal of Economics* 116 (1): 55–79. DOI: https://doi.org/10.1162/003355301556347. [136]

Dato, Simon, Andreas Grunewald, Daniel Müller, and Philipp Strack. 2017. "Expectation-Based Loss Aversion and Strategic Interaction." *Games and Economic Behavior* 104: 681–705. DOI: https://doi.org/10.1016/j.geb.2017.06.010. [123, 134–136, 138, 141, 145, 146]

De la Fuente, Angel. 2000. *Mathematical Methods and Models for Economists*. Cambridge University Press. [139, 150, 154]

de la Rosa, Leonidas Enrique. 2011. "Overconfidence and Moral Hazard." *Games and Economic Behavior* 73 (2): 429–51. DOI: https://doi.org/10.1016/j.geb.2011.04.001. [123, 137]

Enke, Benjamin, and Florian Zimmermann. 2017. "Correlation Neglect in Belief Formation." *Review of Economic Studies* 86 (1): 313–32. DOI: https://doi.org/10.1093/restud/rdx081. [123, 137]

Esponda, Ignacio. 2008. "Behavioral Equilibrium in Economies with Adverse Selection." *American Economic Review* 98 (4): 1269–91. DOI: https://doi.org/10.1257/aer.98.4.1269. [135]

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk–Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models." *Econometrica* 84 (3): 1093–130. DOI: https://doi.org/10.3982/ECTA12609. [123, 126, 132, 135, 138, 141, 151, 152, 155]

Eyster, Erik. 2019. "Chapter 3 - Errors in Strategic Reasoning." In *Handbook of Behavioral Economics - Foundations and Applications 2*. Edited by B. Douglas Bernheim, Stefano DellaVigna,

and David Laibson. Vol. 2, Handbook of Behavioral Economics: Applications and Foundations 1. North-Holland, 187–259. DOI: https://doi.org/10.1016/bs.hesbe.2018.11.003. [123]

Eyster, Erik, and Matthew Rabin. 2005. "Cursed Equilibrium." *Econometrica* 73 (5): 1623–72. DOI: https://doi.org/10.1111/j.1468-0262.2005.00631.x. [123, 131–133, 135, 138, 141, 142]

Feinberg, E. A., P. O. Kasyanov, and N. V. Zadoianchuk. 2014. "Fatou's Lemma for Weakly Converging Probabilities." *Theory of Probability & Its Applications* 58 (4): 683–89. DOI: https://doi.org/10.1137/S0040585X97986850. [140, 145, 148, 154, 155]

Fudenberg, Drew, and David K. Levine. 2006. "A Dual-Self Model of Impulse Control." *American Economic Review* 96 (5): 1449–76. DOI: https://doi.org/10.1257/aer.96.5.1449. [121]

Gabaix, Xavier. 2012. "Game Theory with Sparsity-Based Bounded Rationality." Working paper. URL: https://cutt.ly/FhRnTCf. [137]

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 129 (4): 1661–710. DOI: https://doi.org/10.1093/qje/qju024. [137]

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological games and sequential rationality." *Games and Economic Behavior* 1 (1): 60–79. DOI: https://doi.org/10.1016/0899-8256(89)90005-5. [123, 134–136, 138, 141, 149, 150]

Glicksberg, I. L. 1952. "A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points." *Proceedings of the American Mathematical Society* 3 (1): 170–74. DOI: https://doi.org/10.2307/2032478. [139]

Jehiel, Philippe. 2005. "Analogy-based expectation equilibrium." *Journal of Economic Theory* 123 (2): 81–104. DOI: https://doi.org/10.1016/j.jet.2003.12.003. [123, 133]

Kahneman, Daniel, and Amos Tversky. 1973. "On the Psychology of Prediction." *Psychological Review* 80 (4): 237–51. DOI: https://doi.org/10.1037/h0034747. [134]

Kőszegi, Botond, and Matthew Rabin. 2007. "Reference-Dependent Risk Attitudes." *American Economic Review* 97 (4): 1047–73. DOI: https://doi.org/10.1257/aer.97.4.1047. [129, 145]

Masatlioglu, Yusufcan, and Collin Raymond. 2016. "A Behavioral Analysis of Stochastic Reference Dependence." *American Economic Review* 106 (9): 2760–82. DOI: https://doi.org/10.1257/aer.20140973. [129]

Milgrom, Paul R., and Robert J. Weber. 1985. "Distributional Strategies for Games with Incomplete Information." *Mathematics of Operations Research* 10 (4): 619–32. DOI: https://doi.org/10.1287/moor.10.4.619. [127, 130, 140–142]

Munkres, James R. 2000. *Topology*. Prentice Hall. [150]

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review* 89 (1): 103–24. DOI: https://doi.org/10.1257/aer.89.1.103. [121]

Parthasarathy, Kalyanapuram Rangachari. 1967. *Probability Measures on Metric Spaces*. Vol. 352, American Mathematical Society. [139, 150]

Phillips, Lawrence D, and Ward Edwards. 1966. "Conservatism in a simple probability inference task." *Journal of Experimental Psychology* 72 (3): 346–54. DOI: https://doi.org/10.1037/h0023653. [134]

Rabin, Matthew. 2002. "Inference by Believers in the Law of Small Numbers." *Quarterly Journal of Economics* 117 (3): 775–816. DOI: https://doi.org/10.1162/003355302760193896. [123, 134]

Rabin, Matthew, and Joel L. Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics* 114 (1): 37–82. DOI: https://doi.org/10.1162/003355399555945. [134]

Rabin, Matthew, and Dimitri Vayanos. 2010. "The Gambler's and Hot-Hand Fallacies: Theory and Applications." *Review of Economic Studies* 77 (2): 730–78. DOI: https://doi.org/10.1111/j.1467-937X.2009.00582.x. [123, 134]

Schwardmann, Peter, Egon Tripodi, and Joel van der Weele. 2019. "Self-Persuasion: Evidence from Field Experiments at Two International Debating Competitions." *Cesifo Working Paper Series*, (7946): URL: https://cutt.ly/AhRWeAV. [121, 122]

Schwardmann, Peter, and Joël van der Weele. 2019. "Deception and Self-Deception." *Nature Human Behaviour* 3 (10): 1055–61. DOI: https://doi.org/10.1038/s41562-019-0666-7. [121, 122]