

ESSAYS ON THE
PSYCHOLOGICAL AND SOCIAL
DETERMINANTS OF ECONOMIC BEHAVIOR

Kumulative Dissertation

UNIVERSITÄT HAMBURG
FAKULTÄT FÜR WIRTSCHAFTS- UND SOZIALWISSENSCHAFTEN

Dissertation

zur Erlangung der Würde der Doktorin/des Doktors der
Wirtschaft- und Sozialwissenschaften

(gemäß der PromO vom 24. August 2010)

vorgelegt von

Arno Apffelstaedt

aus Schleswig

Hamburg, 23. April 2018

Vorsitz: Prof. Dr. Gerd Mühlheuser
Erstgutachter: Prof. Dr. Dr. Lydia Mechtenberg
Zweitgutachter: Prof. Dr. Andreas Lange
Drittgutachter: Prof. Dr. Berno Büchel

Datum der Disputation: 12. September 2018

Acknowledgements

First and foremost, I would like to thank my supervisor Lydia Mechtenberg whose academic guidance, support, and insightful comments were immensely important to this research endeavor. I am also grateful for the encouragement and advice received from other faculty members, in particular Anke Gerber, Andreas Lange, and Gerd Mühlheuser. During my time at the University of Hamburg, many other people devoted their time to help and assist, both academically and morally. I especially want to thank Leonie Baumann for the inspiring and fun hours of discussing economic theory, Claudia Schwirplies for being an excellent sparing partner on econometrics, and Jana Freundt for being an amazing co-author.

I am indebted to Paul Heidhues and Botond Köszegi for their help and advice during several stages of my career, as well as to Andrei Shleifer for giving me the opportunity to visit Harvard, and his valuable time and comments on my research during those months. My research stays at Harvard and the London School of Economics were truly inspiring. At the LSE, Francesco Nava and Balazs Szentes provided valuable support and guidance.

Without the financial support from the German Academic Exchange Service, the University of Hamburg's endowment fund, its Center for a Sustainable University, and the graduate school of the Faculty of Business, Economics and Social Sciences most of my conference travels, research visits, and experimental research would not have been possible. Particular thanks go to the head of the graduate school, Ulf Beckmann, for his sincere dedication to providing the department's PhD students with the courses and funds necessary to achieve their goals.

My final thanks go to my family and close friends. Their encouragement and enthusiasm for my work albeit my regular decision to spend weekends and holidays at the desk in front of my papers is worth more than I can express here. This is particularly true regarding my wonderful partner Kristin. Your patience, assistance, support, and faith in me is incredible. Thank you!

Summary

This dissertation presents essays on two questions that have been receiving constantly increasing attention in economics over the last decades: (1) What is the role of psychological, emotional, social and cognitive factors in economic decisions? (2) How can economics incorporate social phenomena such as social norms and conventions, cultural identities and stereotypes, peer and neighborhood effects into its models?

Chapter 1 presents a game theoretic market model that studies the potential influence of psychological attribute salience on consumer choice and market supply in competitive retail markets. Our essay shows that, in equilibrium, retailers strategically manipulate the attribute salience of their products in order to sell naïve consumers a more profitable product than the consumer intended to buy when entering the store. Depending on parameter values, the retailer either sells a more expensive product of higher quality (“up-selling”) or a cheaper product of lower quality (“down-selling”). In both cases, the retailer exploits comparisons with seemingly irrelevant products (“decoys”) in order to increase the salience of the advantageous attribute (quality or price) of the product it aims to sell. The result holds under perfect retailer competition, is robust to the existence of sophisticated and rational consumers, and resonates with anecdotal evidence on psychological “marketing tricks” of retailers as well as with the experimental literature on so-called “context-effects.”

Chapter 2 explores the phenomenon of “spontaneous discrimination” (as derived by Peski and Szentes, 2013: “Spontaneous Discrimination,” *American Economic Review*, 103(6): 2412–2436). Spontaneous discrimination refers to inefficient equilibria in dynamic matching games that are characterized by the seemingly arbitrary coordination of tolerant individuals on a group norm that generates reputational rewards for group members who restrict their interactions to partners of a certain color. To sustain such a norm, information about the color of immediate as well as historical partners has to be revealed to other members of the group. Only then do the reputational mechanisms bite. Chapter 2 develops a theoretical framework to study incentives for information disclosure and analyzes the circumstances in which individuals themselves reveal the color of partners (self-reports) and those in which observers do so (observer-reports). The essay shows that disclosure incentives depend

on whether the market for partners is competitive. While incentives for disclosure do not exist in the non-competitive environment of the benchmark model, they can be created by extending the model to include competition. Competition results in one group benefiting from the discrimination of the other group. Individuals disclose information strategically to gain access to the group that benefits as well as to exclude others from it. Competition also generates incentives for groups to coordinate on a discriminatory norm in the first place. The model can rationalize the observation that individuals sometimes seek group status through discrimination and stigmatization and that groups frequently call for discriminatory rules against outsiders to secure its members access to profitable partnerships (e.g., jobs).

Chapter 3 presents the results of an online-experiment on the question of whether electoral corruption undermines people's willingness to follow democratically elected rules of conduct. Rules concern the redistribution of income. We implement elections in which 100 participants ballot on whether there should exist a rule that asks for the sharing of private (experimental) income or a rule that asks for the opposite. After the election we observe participants' voluntary compliance with the elected rule. The study compares the number of subjects who comply with the rule after an unbiased election with the number of subjects who comply when, during the election, (1) subjects were asked to pay for their vote, (2) subjects were offered money for voting differently, (3) subjects with low household income were excluded from the ballot. In all three cases the data shows a strong and significant reduction in compliance with rules that ask for redistribution. We find no such effect with regard to compliance with the opposite rule ("don't redistribute"). The result suggests that compliance with prosocial rules is affected to a larger extent by corruption than compliance with antisocial rules. Earlier experiments could already demonstrate pure democracy effects in prosocial behavior, but did not deal with either corruption effects or antisocial rules. The study also examines the psychological mechanisms underlying the observed behavior: Treatment effects seem to be driven by intrinsic concerns about procedural aspects of the electoral mechanism, and are particularly prevalent among individuals who express high value for democratic institutions and low value for bribing and (political) lobbying in the real world.

Zusammenfassung

Diese Dissertation legt Aufsätze zu zwei Fragen vor, die in den letzten Jahrzehnten immer mehr Beachtung in der Volkswirtschaftslehre gefunden haben: (1) Welche Rolle spielen psychologische, emotionale, soziale und kognitive Faktoren in ökonomischen Entscheidungen? (2) Wie kann die Volkswirtschaftslehre soziale Phänomene wie soziale Normen und Konventionen, kulturelle Identitäten und Stereotypen, Peer-Group- und Nachbarschaftseffekte in ihre Modelle einbeziehen?

Kapitel 1 untersucht, mithilfe eines spieltheoretischen Marktmodells, den möglichen Einfluss von psychologischer Attributsalienz auf das Kaufverhalten von Konsumenten und das Angebot von Einzelhändlern in kompetitiven Endkonsumentenmärkten. Der Aufsatz zeigt, dass im Gleichgewicht Händler die Attributsalienz ihrer Produkte strategisch manipulieren, um naiven Konsumenten nach Eintritt in das Geschäft ein profitableres als das vom Konsumenten ursprünglich bevorzugte Produkt zu verkaufen. Je nach Parameterwerten verkauft der Händler entweder ein qualitativ höherwertiges, jedoch teureres Produkt (“up-selling”), oder ein billigeres, jedoch qualitativ minderwertigeres Produkt (“down-selling”). In beiden Fällen nutzt der Händler den Vergleich zu scheinbar irrelevanten Produkten (“Decoys”), um die Salienz des vorteilhaften Attributs (Qualität oder Preis) des zu verkaufenden Produkts zu erhöhen. Das Ergebnis hält im perfektem Wettbewerb, ist robust gegenüber der Existenz nicht-naiver und rationaler Konsumenten, und ist im Einklang mit qualitativer Evidenz zu psychologischen “Marketingtricks” von Einzelhändlern, sowie mit der experimentellen Literatur zu sogenannten “Kontexteffekten.”

Kapitel 2 beschäftigt sich mit dem Phänomen der “spontanen Diskriminierung” (aufbauend auf Peski und Szentes, 2013: “Spontaneous Discrimination,” *American Economic Review*, 103(6): 2412–2436). Das Phänomen bezieht sich auf Gleichgewichte in dynamischen Matching-Spielen, die durch die scheinbar willkürliche Koordination von toleranten Individuen auf eine diskriminierende Gruppennorm gekennzeichnet sind, welche mittels endogener Reputationseffekte die ausschließliche Interaktion mit Partnern einer bestimmten Farbe belohnt. Um eine solche Norm aufrechtzuerhalten, muss die Farbe von unmittelbaren und historischen Partnern anderen Mitgliedern der Gruppe offenbart werden. Nur dann greifen die Reputationsmechanismen.

Kapitel 2 entwickelt einen theoretischen Rahmen, um Anreize für die Offenlegung solcher Information zu untersuchen, und analysiert, unter welchen Umständen Individuen selbst die Farbe ihrer Partner offenlegen (Selbstberichte) und unter welchen Umständen Beobachter dies tun (Beobachterberichte). Der Aufsatz zeigt, dass Offenlegungsanreize davon abhängen, ob der Markt für Partner kompetitiv ist. Während im nicht-kompetitiven Markt des Benchmark-Modells keine Anreize zur Offenlegung existieren, können diese durch eine Erweiterung des Modells um Wettbewerb geschaffen werden. Wettbewerb führt dazu, dass eine Gruppe von der Diskriminierung der anderen Gruppe profitiert. Individuen nutzen die Informationsweitergabe in diesem Fall strategisch, um einerseits selbst Zugang zu der bevorzugten Gruppe zu erhalten, und andererseits, um andere aus dieser Gruppe auszuschließen. Auf Gruppenebene schafft Wettbewerb zudem Anreize, sich von vornherein auf eine diskriminierende Norm zu koordinieren. Das Modell kann die Beobachtung rationalisieren, dass Individuen manchmal versuchen, durch Diskriminierung und Stigmatisierung die Zugehörigkeit zu einer Gruppe zu signalisieren, und dass Gruppen häufig diskriminierende Regeln gegen Außenstehende fordern, um ihren Mitgliedern den Zugang zu profitablen Partnerschaften (z.B. Arbeitsplätzen) zu sichern.

Kapitel 3 präsentiert die Ergebnisse eines Online-Experiment zu der Frage, ob Wahlkorruption die Bereitschaft im Volk untergräbt, demokratisch gewählten Verhaltensregeln zu folgen. Die im Experiment untersuchten Verhaltensregeln betreffen die Umverteilung von Einkommen. Wir implementieren Wahlen, in denen jeweils 100 Teilnehmer abstimmen, ob es eine Verhaltensregel geben soll, die dazu auffordert, privates (experimentelles) Einkommen mit anderen zu teilen, oder ob es eine Regel geben soll, die das Gegenteil fordert. Nach der Wahl beobachten wir die freiwillige Einhaltung der gewählten Regel. Die Studie vergleicht die Anzahl an Personen, die sich nach einer unbeeinflussten Wahl an die Regel halten mit der Anzahl an Personen, die sich an die Regel halten, wenn während der Wahl (1) Teilnehmer dazu aufgefordert wurden, Geld für ihre Stimme zu zahlen, (2) Teilnehmer Geld angeboten bekamen, um ihre Stimme zu ändern, (3) Teilnehmer mit einem geringen Haushaltseinkommen von der Wahl ausgeschlossen wurden. Die Daten zeigen in allen drei Fällen einen starken, signifikanten Rückgang bei der Einhaltung von Regeln, die eine Umverteilung fordern. Die Einhaltung der gegenteiligen Regel ("verteile nicht um") ist von diesem Effekt nicht betroffen. Das Ergebnis deutet darauf hin, dass die Einhaltung von prosozialen Regeln stärker von Korruptionseffekten beeinflusst ist als die Einhaltung von nicht-prosozialen Regeln. Frühere experimentelle Studien konnten

bereits reine Demokratieeffekte bei prosozialem Verhalten nachweisen, beschäftigten sich jedoch weder mit Korruptionseffekten noch mit nicht-prosozialem Verhalten. Die Studie untersucht auch die dem beobachteten Verhalten zugrundeliegenden psychologischen Mechanismen: Die Treatmenteffekte scheinen von intrinsischen Bedenken hinsichtlich der prozeduralen Aspekte des Wahlmechanismus getrieben zu sein und finden sich vor allem bei Individuen, die demokratische Institutionen hoch sowie Bestechungsversuche und (politische) Lobbyarbeit in der realen Welt gering schätzen.

List of Included Essays

Chapter 1: Competition over Context-Sensitive Consumers

Authors: Arno Appfelstaedt and Lydia Mechtenberg

Chapter 2: Reputational Discrimination

Author: Arno Appfelstaedt

Chapter 3: Corrupted Votes and Rule Compliance

Authors: Arno Appfelstaedt and Jana Freundt

Contents

Introduction	1
References	5
1 Competition over Context-Sensitive Consumers	8
1.1 Introduction	8
1.2 A Model	12
1.3 Setting the Stage: Attraction and Fooling	15
1.4 Fooling Naïve Populations	17
1.5 Fooling Mixed Populations	23
1.6 Conclusion	24
References	28
Appendix to Chapter 1	31
2 Reputational Discrimination	47
2.1 Introduction	47
2.2 Benchmark: Spontaneous Discrimination	51
2.3 Endogenous Information Disclosure	58
2.4 Competition for Interactions	63
2.5 Conclusion	80
References	83
3 Corrupted Votes and Rule Compliance	86
3.1 Introduction	86
3.2 Experimental Design	93
3.3 Treatment Effects	102
3.4 Understanding Treatment Effects	109
3.5 Conclusion	116
References	119
Appendix to Chapter 3	124

Introduction

This dissertation presents essays on two questions that have been receiving constantly increasing attention in economics over the last decades: (1) What is the role of psychological, emotional, social and cognitive factors in economic decision making? (2) How can economics incorporate social phenomena such as social norms and conventions, cultural identities and stereotypes, peer and neighborhood effects into its models? The two questions can be read as short definitions of research in *behavioral economics* and *social economics*,¹ both of which received individual Journal of Economic Literature (JEL) classification codes as recently as spring 2017.²

While psychological considerations have influenced economic thought throughout history,³ the inception of the contemporary field of behavioral economics is largely credited to experimental and theoretical research on human decision processes conducted in the 1970s and 1980s, in particular to work by Daniel Kahneman, Richard Thaler and Amos Tversky.⁴ Since then, “behavioral” approaches have permeated all field of economics, covering topics from cigarette consumption (Viscusi, 1990) to central bank policy (Ball, Mankiw and Reis, 2005). With psychological factors becoming widely acknowledged as an important ingredient in positive theories of individual decision making, the field has started to move away from revealing behavioral deviations from the standard model of rational choice to exploring the consequences of those deviations for *aggregate* (welfare-relevant) outcomes as well as to finding ways to consolidate the disparate behavioral phenomena with the help of more fundamental, and thus, comprehensive psychological mechanisms. These developments

¹I use the term “social economics” in the tradition of Gary Becker and Kevin Murphy (Becker and Murphy, 2000), following the definition in the *Handbook of Social Economics* (Benhabib, Bisin and Jackson, 2011, p. xvii).

²JEL codes D90/D91 and B55, respectively. See <https://www.aeaweb.org/econlit/jelCodes.php> for the entire list of current codes.

³See, for example, Thaler (2016) for quotes of Adam Smith on loss aversion (“Pain ... is, in almost all cases, a more pungent sensation than the opposite and correspondent pleasure,” Smith, 1759, pp. 176–171), and present bias (“The pleasure which we are to enjoy ten years hence, interests us so little in comparison with that which we may enjoy today,” Smith, 1759, p. 273), as well as Vilfredo Pareto on the role of psychology in economics (“The foundation of political economy, and, in general of every social science, is evidently psychology,” Pareto, 1906, p. 21).

⁴Important seminal studies include Tversky and Kahneman (1974); Kahneman and Tversky (1979); Thaler (1980); Kahneman, Knetsch and Thaler (1986).

form the background for chapter 1 of this dissertation, where we study the consequences of “local thinking”—a behavioral choice theory based on fundamental rules of visual perception—for aggregate market outcomes.

Social economics, in comparison, is at the moment still rather an umbrella term for research that studies how the social environment shapes people’s choices and behaviors than an established field. The term originates from a collection of essays by Gary Becker and Kevin Murphy (Becker and Murphy, 2000) stressing the importance of capturing “culture, norms, and social structure” in economic models. The recent *Handbook of Social Economics* defines the term, in the spirit of Becker and Murphy, as “the study, with the *methods of economics*, of social phenomena in which aggregates affect individual choices. Such phenomena include, just to mention a few, social norms and conventions, cultural identities and stereotypes, peer and neighborhood effects” (Benhabib, Bisin and Jackson, 2011, p. xvii).⁵ While such phenomena can be (and frequently are) approached “behaviorally,” for example, by enriching the utility function with social parameters,⁶ the methodological toolbox of social economics is considerably richer. “Neo-classical” approaches, including evolutionary games, dynamic games, and games on social networks, have proven particularly useful to model the endogenous emergence of social preferences and norms.⁷ Chapter 2 presents a study of this form to model the emergence of discriminatory social norms in tolerant societies.

Finally, chapter 3 combines elements of both, behavioral and social economics. In comparison to the economic theory presented in previous chapters, the research presented here is empirical: We study—using an online experiment with international subjects—how private giving decisions are affected by the democratic election of a voluntary code of conduct, and how the willingness to follow the code is affected by experiencing corruption during the election. The chapter is an essay in *behavioral economics* because it supports a theory of rule compliance that acknowledges psychological factors such as whether the rule has been selected in a fair and democratic manner. It is an essay in *social economics* as it studies how “aggregates [in our case voting outcomes] affect individual choices.” (Benhabib, Bisin and Jackson,

⁵As such, it should be distinguished from *Economic sociology*, “which may be thought of as the study, with the methods of sociology, of economic phenomena, e.g., markets” (ibid, p. xvii).

⁶See, for example, Becker (1957) on preferences for discrimination, Akerlof and Kranton (2000) on preferences for social identity, and Bénabou and Tirole (2012) on (intrinsic) preferences for norm compliance.

⁷See the chapters by Postlewaite (2011), Burke and Young (2011) and Bloch and Duttar (2011) in the *Handbook of Social Economics*.

2011, p. xvii).

Below, I shortly outline the contribution of each chapter in more detail.

Chapter 1 contributes to the literature on *Behavioral Industrial Organization*, which studies the question of whether behavioral deviations from rational choice make consumers susceptible to exploitation by profit-maximizing firms.⁸ We present a game theoretic market model that studies the potential influence of psychological attribute salience (see, e.g. Bordalo, Gennaioli and Shleifer, 2013) on consumer choice and market supply in competitive retail markets. We show that, in equilibrium, retailers strategically manipulate the attribute salience of their products in order to sell naïve consumers a more profitable product than the consumer intended to buy when entering the store. Depending on parameter values, the retailer either sells a more expensive product of higher quality (“up-selling”) or a cheaper product of lower quality (“down-selling”). In both cases, the retailer exploits comparisons with seemingly irrelevant products (“decoys”) in order to increase the salience of the advantageous attribute (quality or price) of the product it aims to sell. The result holds under perfect retailer competition, is robust to the existence of sophisticated and rational consumers, and resonates with anecdotal evidence on psychological “marketing tricks” of retailers as well as with the experimental literature on so-called “context-effects.”

Chapter 2 contributes to the understanding of the social phenomenon of discrimination. In prevalent models, the avoidance of productive interactions with individuals of another color is explained by immediate payoff effects for the decision maker.⁹ In chapter 2, I explore a different possibility, which is that discrimination arises from reputational (that is, *intertemporal*) concerns. In particular, I explore the concept of “spontaneous discrimination” (Peski and Szentes, 2013). Spontaneous discrimination refers to inefficient equilibria in dynamic matching games that are characterized

⁸The literature, as summarized by Spiegler (2011), studies, for example, whether observed pricing, marketing and product differentiation strategies can be explained as equilibrium responses to bounded rationality, and—with regard to market regulation and consumer protection policies—whether market forces (a.k.a. competition) alone can protect consumers from exploitation.

⁹Most existing models use either a “taste-based” (Becker, 1957) or “statistical” (Arrow, 1973; Phelps, 1972) explanation. In models of *taste-based* discrimination, individuals have an inherent preference for interactions with agents of a given (typically their own) color. Models of *statistical* discrimination, on the other hand, assume that agents of one color statistically differ in some payoff-relevant characteristic from the other color. For instance, agents of one color might have higher productivity or crime rates *on average*. Because color can serve as an informative signal of this payoff-relevant factor, even *per-se* tolerant individuals may then discriminate on the margin.

by the seemingly arbitrary coordination of tolerant individuals on a group norm that generates reputational rewards for group members who restrict their interactions to partners of a certain color. To sustain such a norm, information about the color of immediate as well as historical partners has to be revealed to other members of the group. Only then do the reputational mechanisms bite. I develop a theoretical framework to study incentives for information disclosure and analyze the circumstances in which individuals themselves reveal the color of partners (self-reports) and those in which observers do so (observer-reports). The essay shows that disclosure incentives depend on whether the market for partners is competitive. While incentives for disclosure do not exist in the non-competitive environment of the benchmark model, they can be created by extending the model to include competition. Competition results in one group benefiting from the discrimination of the other group. Individuals disclose information strategically to gain access to the group that benefits as well as to exclude others from it. Competition also generates incentives for groups to coordinate on a discriminatory norm in the first place. The model can rationalize the observation that individuals sometimes seek group status through discrimination and stigmatization and that groups frequently call for discriminatory rules against outsiders to secure its members access to profitable partnerships (e.g., jobs).

Chapter 3 contributes to answering the question of how political institutions may interact with economic behavior. The essay presents the results of an online-experiment on the question of whether electoral corruption undermines people’s willingness to follow democratically elected rules of conduct. Rules concern the redistribution of income. We implement elections in which 100 participants ballot on whether there should exist a rule that asks for the sharing of private (experimental) income or a rule that asks for the opposite. After the election we observe participants’ voluntary compliance with the elected rule. The study compares the number of subjects who comply with the rule after an unbiased election with the number of subjects who comply when, during the election, (1) subjects were asked to pay for their vote, (2) subjects were offered money for voting differently, (3) subjects with low household income were excluded from the ballot. In all three cases the data shows a strong and significant reduction in compliance with rules that ask for redistribution. We find no such effect with regard to compliance with the opposite rule (“don’t redistribute”). The result suggest that compliance with prosocial rules is affected to a larger extent by corruption than compliance with antisocial rules. Earlier experiments could already demonstrate pure democracy effects in prosocial behavior (see,

e.g. Dal Bó, Foster and Putterman, 2010), but did not deal with either corruption effects or antisocial rules. The study also examines the psychological mechanisms underlying the observed behavior: Treatment effects seem to be driven by intrinsic concerns about procedural aspects of the electoral mechanism, and are particularly prevalent among individuals who express high value for democratic institutions and low value for bribing and (political) lobbying in the real world.

Jointly, the three chapters highlight the important role psychological and social factors can play in economic decisions. Chapter 1 shows how the decisions of rational, profit-maximizing agents (firms) may depend on whether the behavior of other agents (consumers) is influenced by psychological factors. Chapter 2 gives one example of how economics can incorporate social phenomena into its models. Set out to explain one particular phenomenon (discrimination in tolerant societies), the model ultimately touches on many (for example, stigmatization, social image, and group identities). Chapter 3 suggests that there is a psychological component in how people react to corruption in elections, raising the important question for future research of how this phenomenon may be captured in economic models.

References

- Akerlof, George A., and Rachel E. Kranton.** 2000. “Economics and Identity.” *The Quarterly Journal of Economics*, 115(3): 715–753.
- Arrow, Kenneth J.** 1973. “The Theory of Discrimination.” In *Discrimination in Labor Markets*, ed. Orley C. Ashenfelter and Albert Everett Rees, 3–33. Princeton: Princeton University Press.
- Ball, Laurence, Gregory N. Mankiw, and Ricardo Reis.** 2005. “Monetary Policy for Inattentive Economies.” *Journal of Monetary Economics*, 52(4): 703–725.
- Becker, Gary S.** 1957. *The Economics of Discrimination*. Chicago: The University of Chicago Press.
- Becker, Gary S., and Kevin M. Murphy.** 2000. *Social Economics: Market Behavior in a Social Environment*. Cambridge, Mass.: Harvard University Press.

- Bénabou, Roland, and Jean Tirole.** 2012. “Laws and Norms.” *IZA Discussion Paper No. 6290*.
- Benhabib, Jess, Alberto Bisin, and Matthew O. Jackson,** ed. 2011. *Handbook of Social Economics*. Vol. 1A, Amsterdam: North Holland.
- Bloch, Francis, and Bhaskar Duttar.** 2011. “Formation of Networks and Coalitions.” In *Handbook of Social Economics*. Vol. 1A, ed. Jess Benhabib, Alberto Bisin and Matthew O. Jackson, 729–779. Amsterdam: North Holland.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer.** 2013. “Salience and Consumer Choice.” *Journal of Political Economy*, 121(5): 803–843.
- Burke, Mary A., and H. Peyton Young.** 2011. “Social Norms.” In *Handbook of Social Economics*. Vol. 1A, ed. Jess Benhabib, Alberto Bisin and Matthew O. Jackson, 311–338. Amsterdam: North Holland.
- Dal Bó, Pedro, Andrew Foster, and Louis Putterman.** 2010. “Institutions and Behavior: Experimental Evidence on the Effects of Democracy.” *American Economic Review*, 100(5): 2205–2229.
- Kahneman, Daniel, and Amos Tversky.** 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica*, 47(2): 263–291.
- Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler.** 1986. “Fairness as a Constraint on Profit Seeking: Entitlements in the Market.” *American Economic Review*, 76(Sep., 1986): 728–741.
- Pareto, Vilfredo.** 1906. *Manual of Political Economy*. Translated and annotated reprint, 2014, ed. A. Montesano, A. Zanni, L. Bruni, J. S. Chipman, and M. McLure. Oxford: Oxford University Press.
- Peski, Marcin, and Balasz Szentes.** 2013. “Spontaneous Discrimination.” *American Economic Review*, 103(6): 2412–2436.
- Phelps, Edmund S.** 1972. “The Statistical Theory of Racism and Sexism.” *American Economic Review*, 62(4): 659–661.
- Postlewaite, Andrew.** 2011. “Social Norms and Preferences.” In *Handbook of Social Economics*. Vol. 1A, ed. Jess Benhabib, Alberto Bisin and Matthew O. Jackson, 31–67. Amsterdam: North Holland.

- Smith, Adam.** 1759. *The Theory of Moral Sentiments*. Reprint, 1981, ed. D. D. Raphael and A. L. Macfie. Indianapolis: Liberty Classics.
- Spiegler, Ran.** 2011. *Bounded Rationality and Industrial Organization*. Oxford: Oxford University Press.
- Thaler, Richard H.** 1980. "Toward a Positive Theory of Consumer Choice." *Journal of Economic Behavior & Organization*, 1(1): 39–60.
- Thaler, Richard H.** 2016. "Behavioral Economics: Past, Present, and Future." *American Economic Review*, 106(7): 1577–1600.
- Tversky, Amos, and Daniel Kahneman.** 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131.
- Viscusi, W. Kip.** 1990. "Do Smokers Underestimate Risks?" *Journal of Political Economy*, 98(6): 1253–1269.

Chapter 1

Competition over Context-Sensitive Consumers

Authors: Arno Apffelstaedt and Lydia Mechtenberg

Abstract: We study a model of a competitive retail market in which consumer preferences are sensitive to local salience effects at the store (modeled by nesting recent theories of Bordalo, Gennaioli and Shleifer, 2013; Kőszegi and Szeidl, 2013; Bushong, Rabin and Schwartzstein, 2016). Our main result connects anecdotal evidence on retailer marketing tricks with the experimental literature on context-effects. In equilibrium, retailers use a “fooling strategy”: They attract naïve consumers to their store with a competitive bait product, but then use decoy effects to induce a switch to more profitable alternatives featuring higher price (up-selling) or lower quality (down-selling).

Keywords: Choice Context, Salience, Up-Selling, Down-Selling, Decoys

JEL Codes: D91, D11, D41

1.1 Introduction

Many people are *local thinkers*: We perceive \$10 for a given bottle of wine to be expensive when accompanied by cheaper alternatives (say, at a discount store), but cheap at an exclusive liquor store where alternatives cost \$20 on average. A range of promising theories have recently emerged to model such behavior, reflecting the observation that consumers judge alternatives relative to the immediate environment in which they are presented, among them the theories of *Salience* (Bordalo, Gennaioli and Shleifer, 2013), *Focusing* (Kőszegi and Szeidl, 2013), and *Relative Thinking* (Bushong, Rabin and Schwartzstein, 2016).

This essay studies an important yet unexplored consequence of local thinking

in markets, which is that consumer preferences when planning a purchase (say, at home) may be different from preferences when ultimately making the purchase (at the store). Consider yourself planning the purchase of that bottle of wine at home. Are you aware that you are willing to spend more money for a similar bottle at the liquor store than at the discount supermarket? We show—by modeling a competitive retail market with local thinkers—that if consumers under-estimate (even just marginally) the effect of context on their choice, sellers will exploit this bias by designing choice environments that drive a wedge between the preferences inside and outside of the store. Sellers use this wedge to compete for the consumer with an unprofitable *attraction product*, knowing that the choice environment at the store will ultimately make her prefer a more profitable *target product*. When preferences at the store follow a salience characterization along the lines of Bordalo, Gennaioli and Shleifer (2013), Köszegi and Szeidl (2013) or Bushong, Rabin and Schwartzstein (2016), sellers generate preference distortions using *decoys*: They present product lines that contain a seemingly irrelevant third alternative, which—for a local thinker—makes the target stand out in relative value at the store. Equilibrium product lines are then remarkably similar to choice sets that have been experimentally shown to induce preference reversals (see, e.g., Huber, Payne and Puto, 1982; Simonson, 1989).

The marketing strategies we predict bear strong resemblance to the retail market phenomena known as *up-selling* and *down-selling*—sellers inducing switches to more profitable products using a smart presentation of options at the final point of purchase. Most consumers come across such attempts on a regular (if not even daily) basis.¹ Marketing blogs are abundant with “tricks” on how to design the product line and with hints that consumer naïveté about preference changes lies at the core of the phenomena. They describe up- and down-selling as “getting the consumer to make a higher cost purchase than he or she *originally planned*”, selling “a product that is more expensive than the one *they initially came to buy*” or something more profitable “than the original product they *intended to buy*”.²

In our model, both up-selling equilibria and down-selling equilibria emerge en-

¹Ellison and Ellison (2009) present evidence of such strategies in the online retail market for computer parts. See, also, Max Nisen on “Super cheap airline fares lures in lots of fliers, but most shell out to upgrade” (Quartz, 16th July 2015, retrieved from <https://qz.com/456017>, accessed February 23, 2017) and, for a range of anecdotal examples, <https://econsultancy.com/blog/66879-10-powerful-examples-of-upselling-online/> (accessed February 22, 2017).

²See www.forbes.com/sites/neilpatel/2015/12/21/how-to-upsell-any-customer, <http://www.brainsins.com/en/blog/upselling-increasing-profits/1488>, and <https://www.123-reg.co.uk/blog/ecommerce/how-to-increase-revenue-with-up-selling-and-cross-selling/> (all three have emphasis added and were accessed February 23, 2017).

dogenously. In an up-selling equilibrium, consumers expect to purchase a cheap, low quality product when entering a store, but then shell out to upgrade to a product of higher quality and higher price. In a down-selling equilibrium, retailers sell products of lower quality (and lower cost), while initially attracting the consumer with a product of very high quality. The unique type of marketing strategy that emerges in equilibrium depends on the salience characterization we use as well as on preference and cost parameters. Down-selling regimes tend to emerge when consumers are in principle willing to spend a large amount of money on the product and the cost of producing quality are high. This finding resonates well with the anecdotal evidence on down-selling, which mainly associates retailers of up-scale, luxury products with the phenomenon.³

While rational and sophisticated consumers are not prone to the up- and down-selling strategies that sellers employ in our model, we also show that their presence does not help naïves. We predict that the market reacts to sophisticated consumers by providing separate, non-distortionary stores that naïves do not enter. Rational consumers, on the other hand, enter the distortionary stores which are designed to up- or down-sell naïves and re-exploit them by purchasing the non-profitable attraction product. However, this does not stop sellers from using this practice. Instead, they increase the prices on naïves in order to substitute for the losses made on rational consumers.

Theoretical contributions dealing with context-sensitive consumer preferences in markets are rare. Kamenica (2008) shows that, given that there is also uncertainty about the production cost, a monopolist may be able to change the quality perception of rational, uninformed consumers by adding decoy products to the product line. While this is an important result that sheds new light on the importance of consumer inference, it is definitely not the end of the story. Context-effects have been found in experimental settings with no explanatory room for inference, see, e.g., Herne (1999), Ariely, Loewenstein and Prelec (2003), Mazar, Köszegi and Ariely (2014) and Jahedi (2011). Moreover, the conjecture that context-sensitive shopping behavior is largely irrational seems corroborated by the extensive online discussion of context- and salience-related marketing techniques that all seem to “manipulate” or “trick” consumers into purchase decisions.

³Christina Binkley makes a convincing case for this marketing strategy to be wide-spread in the high-fashion industry in her aptly named article “The Psychology of the \$14,000 Handbag: How Luxury Brands Alter Shoppers’ Price Perceptions; Buying a Keychain Instead” (The Wall Street Journal, 9th August 2007, retrieved from <https://www.wsj.com/articles/SB118662048221792463>, accessed February 23, 2017).

Earlier literature in behavioral economics has made the point that context matters, but has not formally studied its strategic role in competitive markets.⁴ Instead, it has offered theories that are able to explain and model context-dependent preferences. Our model is sufficiently general to encompass these theories, and we produce results for three prominent ones (Bordalo, Gennaioli and Shleifer, 2013; Kőszegi and Szeidl, 2013; Bushong, Rabin and Schwartzstein, 2016) in this essay. We highlight a hitherto unstudied strategic use of context that only exists in competitive markets: Designing choice environments that drive a wedge between consumer preferences in the moment of competition with other firms and preferences in the moment of purchase. It is this particular exploitation of naïve context-sensitivity that generates product lines with three distinct products for just one type of consumer: a false competitor (the attraction product), a target, and a decoy. Such choice sets have inspired early experimental research on context effects (see, in particular, Huber, Payne and Puto, 1982), and have been used as rationale to offer theories of context-dependent consumer choice (most recently by Bordalo, Gennaioli and Shleifer 2013 and Bushong, Rabin and Schwartzstein 2016). To our knowledge, we are the first to provide a model that predicts their existence in markets.

There are other papers in the literature on competition over biased consumers that like ours feature a two-phase choice procedure by which consumers first select a firm and then a product. However, they do not allow local choice environments to affect consumer preferences. Some of these papers relate to ours by the idea that marketing devices or frames play a strategic role when attracting consumers (Eliaz and Spiegler 2011*a*, Eliaz and Spiegler 2011*b*, Piccione and Spiegler 2012), others more technically by the fact that there exists an element of naïve time-inconsistency that firms may try to exploit (among others, Gabaix and Laibson 2006, Ellison 2005, DellaVigna and Malmendier 2004, Heidhues and Kőszegi 2010, and Heidhues, Kőszegi and Murooka 2017). Our results are in many regards novel with regard to both of these streams. A more detailed discussion of our contribution to this literature is relegated to the conclusion.

The remainder of the chapter is organized as follows. We introduce a formal model in the next section. In section 1.3 we derive the equilibrium for rational and sophisticated populations. Section 1.4 derives the equilibrium for naïve consumer populations. Section 1.5 proves that the fooling of naïves persists (and might even

⁴A notable exception is Bordalo, Gennaioli and Shleifer (2016) who, however, do not study the possibility that preferences may change after selecting a seller, which is the assumption lying at the core of our model.

worsen) in consumer populations that also contain sophisticated and rational agents. Section 3.5 concludes with a discussion of our results. All proofs are in the appendix to this chapter.

1.2 A Model

A unit mass of consumers has demand for a good that can be differentiated in quality $q \in \mathbb{R}$ and price $p \in \mathbb{R}$, where quality and price are both measured in monetary units. There is a minimum quality $q_{\min} > 0$ and a maximum price $p_{\max} > 0$ agents are willing to accept and pay, respectively. Each consumer demands one good. There is a large number K of firms in the market. Each firm k owns a store. To purchase from firm k , a consumer has to enter its store. At the store, the firm can offer any menu of products J^k . Each product $j \in J^k$ implements the good at some level of quality $q_j \in \mathbb{R}$ and price $p_j \in \mathbb{R}$. The set $M^k = ((q_j, p_j))_{j \in J^k}$ is called the *product line* of firm k . Instead of entering a store and purchasing a product, consumers can select the outside option of no purchase. The sequence of events is illustrated in Figure 1.1 below.

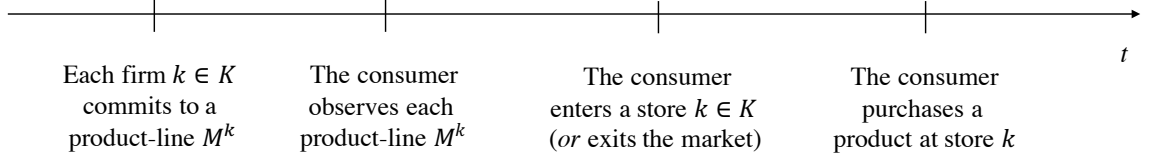


Figure 1.1: Sequence of Events

1.2.1 Product Choice at the Store

Consumers value product j at store k with the *local* surplus function

$$u_j^k(\beta) = \begin{cases} \beta q_j - p_j & \text{if } \theta_j^k = Q \text{ (quality } q_j \text{ is salient at store } k) \\ q_j - \beta p_j & \text{if } \theta_j^k = P \text{ (price } p_j \text{ is salient at store } k) \\ q_j - p_j & \text{if } \theta_j^k = N \text{ (neither is salient at store } k) \end{cases} \quad (1.1)$$

and $\beta \geq 1$. If $\beta > 1$, consumers are sensitive to local salience effects. We call these consumers *local thinkers*. The case of $\beta = 1$ nests the rational consumer. Salience at store k follows one of the following three models:

Assumption BGS (Bordalo, Gennaioli and Shleifer (2013)). “An attribute is salient for a good when it stands out among the good’s attribute relative to that attribute’s average level in the choice set.”⁵

$$\theta_j^k = \begin{cases} Q & \text{if } \sigma(q_j, \bar{q}^k) > \sigma(p_j, \bar{p}^k) \\ P & \text{if } \sigma(q_j, \bar{q}^k) < \sigma(p_j, \bar{p}^k) \\ N & \text{otherwise} \end{cases}$$

where \bar{z}^k is the average level of attribute $z \in \{q, p\}$ at store k and $\sigma(\cdot, \cdot)$ is a symmetric and continuous function that satisfies ordering and homogeneity of degree zero,⁶ for example, $\sigma(z_j, \bar{z}^k) = \frac{z_j - \bar{z}^k}{z_j + \bar{z}^k}$.

Assumption KS (Kőszegi and Szeidl (2013)). “A person focuses more on, and hence overweights, attributes in which her options differ more.”⁷

$$\theta_j^k = \begin{cases} Q & \text{if } (\max_{j \in J^k} q_j - \min_{j \in J^k} q_j) > (\max_{j \in J^k} p_j - \min_{j \in J^k} p_j) \\ P & \text{if } (\max_{j \in J^k} q_j - \min_{j \in J^k} q_j) < (\max_{j \in J^k} p_j - \min_{j \in J^k} p_j) \\ N & \text{otherwise} \end{cases}$$

Assumption BRS (Bushong, Rabin and Schwartzstein (2016)). “Fixed differences loom smaller when compared to large differences.”⁸

$$\theta_j^k = \begin{cases} Q & \text{if } (\max_{j \in J^k} q_j - \min_{j \in J^k} q_j) \cdot \beta < (\max_{j \in J^k} p_j - \min_{j \in J^k} p_j) \\ P & \text{if } (\max_{j \in J^k} q_j - \min_{j \in J^k} q_j) > (\max_{j \in J^k} p_j - \min_{j \in J^k} p_j) \cdot \beta \\ N & \text{otherwise} \end{cases}$$

where $\beta \geq 1$ according to Eq. 1.1.⁹

⁵Cited from the abstract of Bordalo, Gennaioli and Shleifer (2013). The implementation is based on Definition 1 and Assumption 1 in the same paper.

⁶(1) Ordering and (2) homogeneity of degree zero are defined as follows: (1) Let $\mu = \text{sgn}(z_k - \bar{z}^k)$. Then, for any $\varepsilon, \varepsilon' \geq 0$ with $\varepsilon + \varepsilon' > 0$, $\sigma(z_j + \mu\varepsilon, \bar{z}^k - \mu\varepsilon') > \sigma(z_j, \bar{z}^k)$. (2) $\sigma(\alpha z_j, \alpha \bar{z}^k) = \sigma(z_j, \bar{z}^k) \forall \alpha > 0$. In order to work with nonpositive arguments in $\sigma(\cdot, \cdot)$, additional properties are required, see Bordalo, Gennaioli and Shleifer (2013). For our analysis it is sufficient to have $\sigma(\cdot, \cdot)$ defined in the positive domain.

⁷Cited from the abstract of Kőszegi and Szeidl (2013). The implementation is a straightforward adaption of Assumption 1 in Kőszegi and Szeidl (2013) to a setup with discrete utility weights.

⁸Cited from the abstract of Bushong, Rabin and Schwartzstein (2016).

⁹The implementation is based on norming assumptions N0-N2 in Bushong, Rabin and

1.2.2 Choice of Store

Consumers choose a store by predicting their purchase at the store and maximizing the *global* surplus function

$$u_j = q_j - p_j. \quad (1.2)$$

The outside option of not entering a store (not purchasing a product) generates surplus $u_0 = 0$.

A consumer's predictions about her choice behavior inside store k depend on her awareness of local salience effects. We allow for different types, modeled via individual point beliefs regarding the size of factor β , $E(\beta) = \tilde{\beta}$. A consumer with point-belief $\tilde{\beta}$ predicts herself to value products at store k with the surplus function $u_j^k(\tilde{\beta})$. A *sophisticated* consumer has correct belief $\tilde{\beta} = \beta$. A *naïve* consumer has point-belief $\tilde{\beta} \in [1, \beta)$. The lower bound $\tilde{\beta} = 1$ identifies a perfectly naïve type, unaware of local salience effects. Beliefs $\tilde{\beta} \in (1, \beta)$ identify *partially* naïve types who underestimate the impact of salience on their choice.

1.2.3 Firms' Choice of Product-Lines

Firms choose product-lines $M^k = ((q_j, p_j))_{j \in J^k}$, $(q_j, p_j) \in \mathbb{R}^2$, so as to maximize individual profit π^k . They have knowledge of consumer surplus functions and of the distribution of consumer naïveté (regarding salience effects) in the market, but cannot observe the naïveté of individual consumers. Firms have symmetric cost functions. When a consumer purchases a good of quality q from firm k , the firm incurs a cost $c(q)$ that we assume is strictly convex increasing in the quality delivered, $c'(q) > 0$, $c''(q) > 0$, and satisfies $c(0) = c'(0) = 0$. These standard Inada conditions imply that for any form of the local surplus function u_j^k (Eq. (1.1)) there exists a unique,

Schwartzstein (2016). To translate N0-N2 to a setup with discrete utility weights, let $w(\cdot)$ denote the weight function that attaches weight $w_z^k \in \{1, \beta\}$ to attribute $z \in \{q, p\}$. N0 assumes that $w(\cdot)$ is a function of the attribute spread, $w(\Delta_z^k)$. N1 assumes that $w(\Delta_z^k)$ is decreasing in the spread. Finally, N2 assumes that $w(\Delta_z^k) \cdot \Delta_z^k$ is increasing. Our implementation hails mainly from N2. Suppose that quality has a higher weight than price, i.e. $w_q^k = \beta$ and $w_p^k = 1$. According to Equation 1.1, $\theta_j^k = Q$ for all products $j \in J^k$. By N1, $w(\Delta_q^k) > w(\Delta_p^k) \Rightarrow \Delta_q^k < \Delta_p^k$. But N2 makes a more restrictive assumption, namely that $w(\Delta_q^k) > w(\Delta_p^k) \wedge \Delta_q^k < \Delta_p^k \Rightarrow w(\Delta_q^k)\Delta_q^k < w(\Delta_p^k)\Delta_p^k \Leftrightarrow \beta\Delta_q^k < \Delta_p^k$. An analogous statement establishes the case of $\theta_j^k = P$.

strictly positive level of quality q^c that is cost-efficient to sell, namely

$$q^c = \begin{cases} q^Q := \arg \max_q [\beta q - c(q)] \Leftrightarrow c'(q^Q) = \beta & \text{if } \theta_j^k = Q \text{ (quality } q_j \text{ is salient)} \\ q^P := \arg \max_q [q - \beta c(q)] \Leftrightarrow c'(q^P) = \frac{1}{\beta} & \text{if } \theta_j^k = P \text{ (price } p_j \text{ is salient)} \\ q^* := \arg \max_q [q - c(q)] \Leftrightarrow c'(q^*) = 1 & \text{if } \theta_j^k = N \text{ (neither is salient)} \end{cases} \quad (1.3)$$

Note that $q^Q > q^* > q^P > 0$. There is a marginal setup cost $\epsilon \rightarrow 0^+$ for each product added to the product line.

1.2.4 Solution Concept

We analyze market supply in the competitive Nash equilibrium, defined by firms playing mutually best responses and $\pi^k = 0$ for all $k \in K$. We concentrate on interior results by demanding that minimum quality q_{\min} is sufficiently low and maximum willingness to pay p_{\max} sufficiently high that consumers do not *per-se* reject buying cost-efficient quality q^c (Eq. (1.3)) at cost. This is true if and only if $q_{\min} \leq q^P$ and $p_{\max} \geq c(q^Q)$, which we assume henceforth. To resolve possible tie breaks, we make two assumptions. First, whenever indifferent, a consumer chooses each surplus maximizing option with positive probability. Second, there exists a smallest monetary unit $\delta > 0$, which we take to be positive but infinitesimally small.¹⁰ This is equivalent to assuming that a firm, when best-responding, can resolve tie breaks in favor of the strictly more profitable product. We will exploit this equivalence when solving the model.

1.3 Setting the Stage: Attraction and Fooling

We begin with a benchmark. How would market supply look like if consumers were *not* sensitive to salience effects at the store? When $\beta = 1$, local preferences at the store coincide with global preference outside the store. The two-step choice of consumers is irrelevant in such a case. Firm incentives collapse to standard Bertrand incentives: A firm offering the highest global surplus in the market wins all consumers. It follows:

¹⁰Formally, let $\delta = \frac{1}{10^z}$ where $z \in \mathbb{Z}$ is an integer. Firms then choose qualities and prices from a discretized set of real numbers $R_z = \{r \in \mathbb{R} | (r \cdot 10^z) \in \mathbb{Z}\}$. In the limit $z \rightarrow \infty$ (i.e., $\delta \rightarrow 0^+$) this set is equal to \mathbb{R} .

Lemma 1.1 (Rational Benchmark). *Consider a rational consumer population ($\beta = 1$). In equilibrium, consumers purchase quality q^* at price equal to marginal cost, $p = c(q^*)$. (Non-empty) product-lines contain a single product, $M^k = ((q^*, c(q^*)))$.*

Next consider consumers who are sensitive to salience ($\beta > 1$) but *sophisticated*. These consumers have preferences that can be influenced by local stimuli at a store. However, being aware of this bias, they perfectly predict their in-store choices ex-ante. Sophisticated consumers enter store k only if the product they will purchase at store k provides at least as high global surplus as any other product they would buy elsewhere: Due to perfect foresight, the choice of sophisticated consumers between firms is *as if* they were not context-sensitive. Competition for such consumers generates the same incentives as competition for rational consumers.

Proposition 1.1 (Sophisticated populations). *Consider a population of sophisticated local thinkers ($\beta > 1$, $\tilde{\beta} = \beta$). Equilibrium market supply is identical to the rational benchmark (Lemma 1.1).*

Things change when consumers are naïve regarding their sensitivity to salience effects: If preferences are distorted at store k , the product a naïve consumer predicts to buy at the store must not necessarily conform to the product she will ultimately prefer to buy. We therefore define:

Definition 1.1 (Attraction Product). *If there exists a unique product $j \in J^k$ that a consumer with point-belief $\tilde{\beta}$ expects to purchase at store k , we call it the attraction product $a^k(\tilde{\beta})$ of firm k .*

Definition 1.2 (Target). *If there exists a unique product $j \in J^k$ that a consumer purchases when entering store k , we call it the target t^k of firm k .*

Naïveté about salience effects lies at the core of their exploitability: It entails the possibility for firms to design product lines that attract the consumer with a product the firm ultimately does not sell. If a firm employs such a strategy, we say that the firm *fools* the consumer:

Definition 1.3 (Fooling). *Firm k fools a local thinker of type $\tilde{\beta}$ if and only if (1) $a^k(\tilde{\beta})$ and t^k exist and (2) $a^k(\tilde{\beta}) \neq t^k$. If firm k fools type $\tilde{\beta}$,*

$$u_{t^k}^k(\beta) \geq u_{a^k(\tilde{\beta})}^k(\beta) \quad (\text{IC})$$

$$u_{t^k}^k(\tilde{\beta}) \leq u_{a^k(\tilde{\beta})}^k(\tilde{\beta}) \quad (\text{PCC})$$

with at least one of the inequalities being strict.

In this definition, condition (IC) is a standard incentive compatibility constraint: At store k , the consumer prefers the target over the attraction product. When considering to enter store k , however, a fooled consumer *falsely* expects that she will prefer the attraction product over the target: This is covered by the *perceived choice* constraint (PCC).

1.4 Fooling Naïve Populations

Consider a naïve consumer with belief $\tilde{\beta} < \beta$. Fooling can be a profitable strategy because it allows the firm to monopolize on a preference shock that the consumer did not expect when entering the store. Profitable fooling requires an adequate design of (1) the characteristics of the target and attraction product and of (2) the preference shock. Local distortions of consumer preferences at store k matter in so far as they affect the salience of the quality and price of the attraction product a^k and the target t^k . The following lemma addresses the question of which pairs of preference-distortions (θ_a^k, θ_t^k) can be profitably exploited by the firm.

Lemma 1.2 (Profitable Fooling). *Assume that a profit-maximizing firm offers a single product of quality $q_j > q_{\min}$ which it sells at price $p_j < p_{\max}$ to a naïve local thinker ($\beta > 1$, $\tilde{\beta} < \beta$). The firm can strictly increase its profit on the consumer by adding a second product j' to the product line, using one product as target t and the other as attraction product $a \neq t$ if and only if*

1. *The quality of both products is salient at the store, $(\theta_a^k, \theta_t^k) = (Q, Q)$, given that the quality and price of the target is higher than that of the attraction product, $q_t > q_a$ and $p_t > p_a$, or*
2. *The price of both products is salient at the store, $(\theta_a^k, \theta_t^k) = (P, P)$, given that the quality and price of the target is lower than that of the attraction product, $q_t < q_a$ and $p_t < p_a$, or*
3. *Salience effects at the store are asymmetric and distort preferences in favor of the target, $(\theta_a^k, \theta_t^k) \in \{(P, Q), (P, N), (N, Q)\}$.*

Whether a firm can profitably fool—and if so, which of the profitable fooling strategies listed in Lemma 1.2 it will use—depends on particulars of the salience

model employed as well as on consumer and cost characteristics. A central difference between the models suggested by Bordalo, Gennaioli and Shleifer (2013), Kőszegi and Szeidl (2013) and Bushong, Rabin and Schwartzstein (2016) concerns the question whether asymmetric salience effects (Lemma 1.2, point 3) are feasible to construct: Under Assumption KS and Assumption BRS salience effects are necessarily symmetric as they depend on the spread of attributes in the choice set: If the quality (price) of product $j \in J^k$ is salient, then, necessarily, the quality (price) of any other product $j' \in J^k$ must also be salient. Under Assumption BGS, however, distortions depend on product-specific values of the salience function $\sigma(z_j, \bar{z}^k)$, potentially generating asymmetric salience effects. Given the quality and price of the target and attraction product, asymmetric salience effects tend to generate larger (and thus, more profitable) preference shocks because they can increase the consumer's valuation of the target by relatively more than her valuation of the attraction product.

We solve for the equilibrium with naïve consumers in two steps: Proposition 1.2 characterizes the equilibrium under the assumption that firms have an *unspecified* technology at hand that lets them choose preference distortions θ_j^k at their store directly. Firms choose this distortion simultaneously when also designing the product line. We consider the case where this technology allows for asymmetric salience effects (working towards a characterization of the equilibrium under Assumption BGS) and the case where it is restricted to symmetric distortions (working towards a characterization of the equilibrium under Assumptions KS or BRS). After discussing the outcome, Proposition 1.3 then characterizes the equilibrium when distortions are endogenous to the product line as assumed by Bordalo, Gennaioli and Shleifer (2013), Kőszegi and Szeidl (2013) and Bushong, Rabin and Schwartzstein (2016)—showing how firms in this case can use the product line to construct the exact same outcome as if they were choosing preference distortions θ_j^k directly.

Proposition 1.2 (Fooling with an unspecified salience technology). *Consider a population of naïve local thinkers ($\beta > 1$, $\tilde{\beta} < \beta$), possibly with heterogenous degrees of naïveté $\tilde{\beta} < \beta$. Assume that firms have access to an unspecified salience technology that allows them to choose preference distortions θ_j^k for products offered at their store, either being restricted to symmetric distortions, $\theta_j^k = \theta_{j'}^k = \theta^k \in \{Q, P, N\}$ if $\{j, j'\} \subseteq J^k$, or being able to choose symmetric and asymmetric distortions, $(\theta_j^k, \theta_{j'}^k) \in \{Q, P, N\}^2$ for any $\{j, j'\} \subseteq J^k$. In equilibrium, firms choose distortions $\theta_j^k \neq N$. All naïve consumers are fooled. (Non-empty) product lines consist of two products: A (unique) attraction product (attracting all consumers with $\tilde{\beta} < \beta$), and*

a target, $M^k = ((q_{a^k}, p_{a^k}), (q_{t^k}, p_{t^k}))$. Naïve consumers are attracted with a product that is priced below marginal cost $p_{a^k} < c(q_{a^k})$, but ultimately purchase a quality- or price-distorted target at a price equal to marginal cost $p_{t^k} = c(q_{t^k})$.

Equilibrium qualities, prices and distortions are identical across firms. To simplify notation, let $(q_t, p_t) := (q_{t^k}, p_{t^k})$, $(q_a, p_a) := (q_{a^k}, p_{a^k})$ and $(\theta_a, \theta_t) := (\theta_{a^k}^k, \theta_{t^k}^k)$.

1. **Symmetric Distortions.** Assume that firms are restricted to symmetric distortions, $\theta_j^k = \theta_{j'}^k = \theta^k \in \{Q, P, N\}$ if $\{j, j'\} \subseteq J^k$. Define

$$\begin{aligned} \nu^{(Q,Q)} &:= [q^Q - c(q^Q)] + (\beta - 1)(q^Q - q_{\min}), \text{ and} \\ \nu^{(P,P)} &:= [q^P - c(q^P)] + (\beta - 1)[p_{\max} - c(q^P)], \end{aligned}$$

where q^Q and q^P are cost-efficient quality levels as defined in the model section, Eq. (1.3).

- a) If $\nu^{(Q,Q)} \geq \nu^{(P,P)}$, then $(\theta_a, \theta_t) = (Q, Q)$. Firms attract naïves with a product of minimal quality $q_a = q_{\min}$, $p_a < c(q_{\min})$, and up-sell to $(q_t, p_t) = (q^Q, c(q^Q))$.
- b) If $\nu^{(Q,Q)} \leq \nu^{(P,P)}$, then $(\theta_a, \theta_t) = (P, P)$. Firms attract naïves with a product of maximal price $p_a = p_{\max}$, $q_a > c^{-1}(p_{\max})$, and down-sell to $(q_t, p_t) = (q^P, c(q^P))$.

2. **Asymmetric Distortions.** Assume that firms can choose symmetric and asymmetric distortions, $(\theta_j^k, \theta_{j'}^k) \in \{Q, P, N\}^2$ for any $\{j, j'\} \subseteq J^k$. Then $(\theta_a, \theta_t) = (P, Q)$. Firms attract naïves with a product of maximal price $p_a = p_{\max}$, $q_a > c^{-1}(p_{\max})$, and down-sell to $(q_t, p_t) = (q^Q, c(q^Q))$.

Firms choose to distort preferences at their store and fool because this yields higher profits than a classical undercutting strategy (Lemma 1.2). Fooling is profitable regardless of the degree of naïveté. Heterogeneity in this degree is irrelevant because the profit maximizing choice of an attraction product $(q_{a^k(\tilde{\beta})}, p_{a^k(\tilde{\beta})})$ and a target (q_{t^k}, p_{t^k}) for a given degree of naïveté $\tilde{\beta}$ fools naïves of any degree. The choice of *which* type of distortion $(\theta_{a^k}^k, \theta_{t^k}^k)$ to use is essentially a choice for the regime that generates the largest (and thus, most profitable) preference shock. When firms have access to a technology that allows for asymmetric salience effects, the preference shock induced by a simultaneous decrease in attraction product value and increase in target value, $(\theta_{a^k}^k, \theta_{t^k}^k) = (P, Q)$, clearly dominates all other choices.

There is no dominant choice when firms are restricted to symmetric distortions: A quality-salient store, $(\theta_{a^k}^k, \theta_{t^k}^k) = (Q, Q)$, tends to generate a more profitable preference shock when quality is cheap to produce ($c(q)$ is flat) and consumers can be attracted by a product of low quality ($q_{a^k} = q_{\min}$ is small). The preference shock is larger in a price-salient store, $(\theta_{a^k}^k, \theta_{t^k}^k) = (P, P)$, on the other hand, if quality is costly to produce ($c(q)$ is steep) and consumers can be attracted by a product with a high price tag ($p_{a^k} = p_{\max}$ is large). This is in line with the idea that down-selling regimes, $(\theta_{a^k}^k, \theta_{t^k}^k) = (P, P)$, tend to emerge in markets for exclusive (for example, high-fashion) products, while up-selling regimes, $(\theta_{a^k}^k, \theta_{t^k}^k) = (Q, Q)$, are common in markets for everyday consumption goods.

We now move to the characterization of the equilibrium when distortions emerge endogenously as a function of the product line—embedding the theories of Saliency (Bordalo, Gennaioli and Shleifer, 2013), Focusing (Kőszegi and Szeidl, 2013) and Relative Thinking (Bushong, Rabin and Schwartzstein, 2016) in our framework.

Proposition 1.3 (Fooling with Saliency, Focusing, and Relative Thinking). *Consider a population of naïve local thinkers ($\beta > 1$, $\tilde{\beta} < \beta$), possibly with heterogeneous degrees of naïveté $\tilde{\beta} < \beta$. Assume that saliency follows Assumption BGS, KS, or BRS. In equilibrium, firms generate distortions $\theta_j^k \neq N$ using the product line. All naïve consumers are fooled. (Non-empty) product lines consist of three products: A (unique) attraction product (attracting all consumers with $\tilde{\beta} < \beta$), a target, and a decoy, $M^k = ((q_{a^k}, p_{a^k}), (q_{t^k}, p_{t^k}), (q_{d^k}, p_{d^k}))$. Qualities, prices and distortions of the target and attraction product are identical to the case where firms choose distortions directly (Proposition 1.2): The symmetric characterization is valid under Assumptions KS (Focusing) and BRS (Relative Thinking); the asymmetric characterization is valid under Assumption BGS (Saliency).*

Intriguingly, a simple manipulation of the product line allows firms to construct fooling regimes *as if* they were choosing distortions directly: A third product d^k that itself is unattractive as an option for the consumer—both, in expectation ($u_{d^k}^k(\tilde{\beta}) < u_a^k(\tilde{\beta})$) and at the moment of purchase ($u_{d^k}^k(\beta) < u_t^k(\beta)$)—can be designed in such a way that it makes the relevant attributes of the target and attraction product salient at the store, inducing the desired preference shock. This finding resonates with experiments demonstrating so-called *decoy-effects* in consumer choice—preference reversals that can be induced by adding seemingly irrelevant alternatives to the choice set (see, for example, Huber, Payne and Puto, 1982; Doyle et al., 1999; Herne, 1999). The possibility for such violations of the IIA property is nested via the

relation of choice set and salience in the models of Bordalo, Gennaioli and Shleifer (2013), Kőszegi and Szeidl (2013) and Bushong, Rabin and Schwartzstein (2016). Our result shows how competitive firms can systematically exploit this possibility to fool naïve consumers into more profitable purchase decisions. Note that the addition of a decoy is not only sufficient but also *necessary* to generate the desired fooling outcome: With just two products spanning the choice set, the theories of Focusing (Kőszegi and Szeidl, 2013) and Relative Thinking (Bushong, Rabin and Schwartzstein, 2016) imply that consumers behave as if they were maximizing an unweighted surplus function, making it impossible to fool consumers. The theory of Salience (Bordalo, Gennaioli and Shleifer, 2013), on the other hand, implies that choice sets containing only two options can generate symmetric, but not asymmetric salience effects.

The position in quality-price space of an adequate *decoy*—able to generate a profitable preference reversal—depends on which salience model we employ, see Figures 1.2 and 1.3. Figure 1.2 depicts equilibrium locations of the decoy under Assumptions KS (“Focusing”: Kőszegi and Szeidl, 2013) and BRS (“Relative Thinking”: Bushong, Rabin and Schwartzstein, 2016). Both specifications imply that salience effects are symmetric, that is, $\theta_j^k = \theta_{j'}^k$ if $\{j, j'\} \in J^k$. There are two cases, see Proposition 1.2: (a) Firms *up-sell*, $q_t > q_a$ and $p_t > p_a$, using a quality-salient product line, $(\theta_a, \theta_t) = (Q, Q)$ (depicted in the left panel of figure 1.2), and (b) Firms *down-sell*, $q_t < q_a$ and $p_t < p_a$, using a price-salient product line, $(\theta_a, \theta_t) = (P, P)$ (depicted in the right panel). To achieve the profit-maximizing distortion without violating incentive compatibility, firms have to add a decoy to the product line that resides within the boundaries of the grey shaded areas in Figure 1. When salience follows Assumption BGS (Bordalo, Gennaioli and Shleifer, 2013), preference distortions can be asymmetric. In this case, having distortion $(\theta_a, \theta_t) = (P, Q)$ is profit-maximizing for the firm. Figure 1.3 illustrates how the firm can construct this distortion with one decoy. The figure depicts the case when, as in equilibrium, $q_a > q_t$ and $p_a > p_t$ (the firm down-sells). The firm can generate distortion $(\theta_a, \theta_t) = (P, Q)$ in this case by constructing a reference point (\bar{q}^k, \bar{p}^k) that is either strictly dominated by the target ($\bar{p}^k = p_t$, while $\bar{q}^k < q_t$) or by the attraction product ($\bar{q}^k = q_a$, while $\bar{p}^k > p_a$). Which of the two constructions is feasible depends on whether the target or the attraction product has a higher quality-to-price ratio (see the left panel and right panel of Figure 1.3, respectively). In both cases, such a reference point can always be constructed—using a single, unattractive decoy—without violating incentive

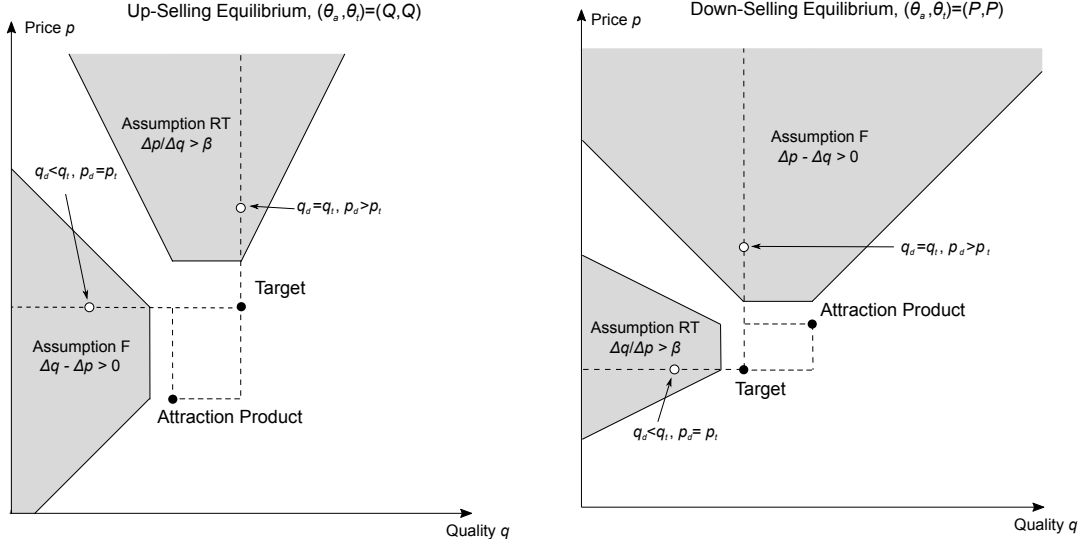


Figure 1.2: Equilibrium choice of decoy (=within shaded areas) under Assumptions KS (Focusing) and BRS (Relative Thinking).

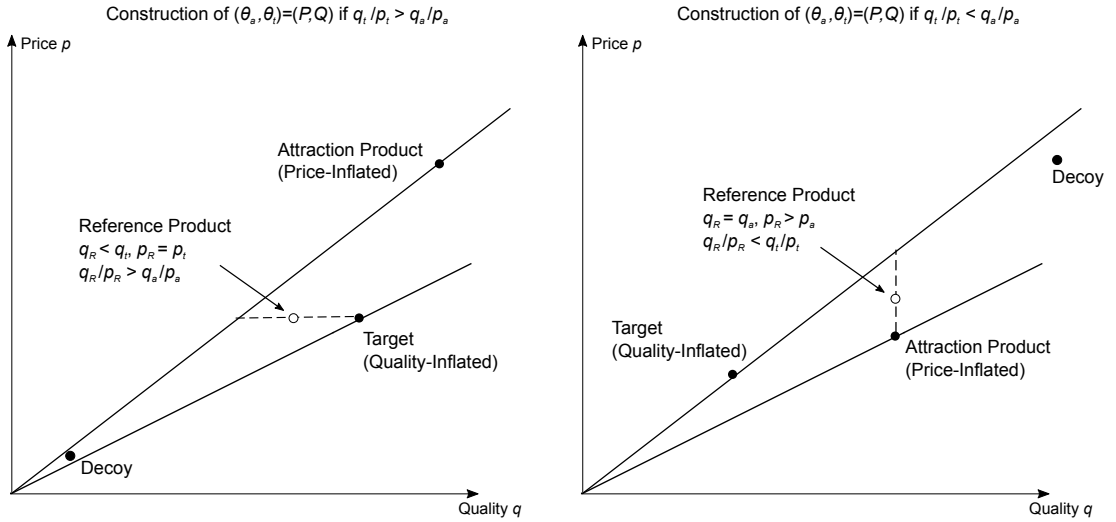


Figure 1.3: Construction of distortion $(\theta_a, \theta_t) = (P, Q)$ (with one decoy) under Assumption BGS.

The construction exploits two central implications of the Saliency framework: (1) If product $j \in J^k$ neither dominates nor is dominated by the reference point, i.e., $(q_j - \bar{q}^k)(p_j - \bar{p}^k) > 0$, then the “advantageous” attribute of product j —higher quality *or* lower price relative to the average—is overweighted if and only if the product has better-than-average quality-to-price ratio, that is, $(q_j / p_j) > (\bar{q}^k / \bar{p}^k)$. (2) If one attribute of product $j \in J^k$ is average while the other is not (e.g., $q_j = \bar{q}^k$, but $p_j \neq \bar{p}^k$), then the latter is overweighted.

compatibility.

1.5 Fooling Mixed Populations

How is the predicted exploitation of naïve consumers affected by the co-existence of sophisticated or rational consumers? We show below that fooling survives in mixed populations. For the following two propositions, let firms either directly choose θ_j^k (with store-wide or product-specific distortions, following the assumptions in Proposition 1.2), or let Assumption BGS, KS, or BRS be satisfied (firms can manipulate θ_j^k indirectly using decoy products).

Proposition 1.4 (Co-Existence of Sophisticated and Naïve Agents). *Consider a population of local thinkers ($\beta > 1$) that contains both, sophisticated agents ($\tilde{\beta} = \beta$) and naïve agents ($\tilde{\beta} < \beta$, of possibly heterogenous degree). In equilibrium, all naïve consumers are fooled, purchasing a quality- or price-distorted target at a price equal to marginal cost, $p = c(q_{tk})$, $q_{tk} \in \{q^Q, q^P\}$ (product supply follows Proposition 1.2 or 1.3, respectively). Sophisticated consumers enter different stores than naïves, purchasing an undistorted product at a price equal to marginal cost, $p = c(q^*)$ (product supply follows the rational benchmark, Lemma 1).*

Proposition 1.5 (Co-Existence of Rational and Naïve Agents). *Consider a consumer population that contains both, rational consumers ($\beta = 1$) and naïve local thinkers ($\beta > 1$, $\tilde{\beta} < \beta$, of possibly heterogenous degree). We concentrate on interior solutions (w.l.o.g., let $p_{\max} \rightarrow \infty$). In equilibrium, all naïve consumers are fooled, purchasing a quality- or price-distorted target at a price above marginal cost, $p > c(q_{tk})$, $q_{tk} \in \{q^Q, q^P\}$. Rational consumers enter the same stores as naïves, but purchase the attraction product at a price below marginal cost, $p < c(q_{ak})$. (See proof for details on product supply.)*

Proposition 1.4 shows that firms react to the introduction of sophisticated consumers with the provision of *additional* non-distortionary stores that allow consumers to self-commit to the ex-ante efficient product (mirroring market supply in the rational benchmark). While all sophisticated consumers sort into these stores, those who under-estimate the effect of salience on their choice expect to be receiving a better deal elsewhere and continue being fooled. Because profitable-to-fool and unprofitable-to-fool consumers are perfectly separated into two types of stores, market supply and exploitation of naïves is completely unaffected by the presence of

more sophisticated agents: Fooling follows our characterization in earlier propositions (Propositions 1.2 and 1.3).

The presence of *rational* consumers ($\beta = 1$), on the other hand, affects the “degree” to which firms are able to fool naïves: Having no commitment problem, rational consumers can enter distortionary firms and re-exploit them by purchasing the (non-profitable) attraction product. However, as Proposition 1.5 shows, the incentive to use context effects to up- or down-sell naïve consumers is *not* lessened. Fooling survives with the result being a trade-off between the profit lost on rational consumers ($p_{a^k} < c(q_{a^k})$) and the profit made by up- or downselling naïves ($p_{t^k} > c(q_{t^k})$). The particular design of product-line and distortion depends on the salience model employed and the share of naïve agents in the population, but is again unique and similar in flavor to our earlier characterizations (see the proof of Proposition 1.5 for detail). Because rational agents gain from the presence of naïves (the bargain of the former being subsidized by the latter), the exploitation of naïves even *increases* compared to the original fooling equilibrium: While the quality they receive ($q_{t^k} \in \{q^Q, q^P\}$) is independent of their share η in the population, they pay a price strictly above cost, $p_{t^k} > c(q_{t^k})$, whenever η is below unity. This is the case even in the limit as $\eta \rightarrow 0$ and rational consumers are provided with the exact same product as in the rational benchmark. This finding shows that fooling may be an important, welfare-relevant phenomenon even when the mass of victims falling prey to such practices is small.

1.6 Conclusion

We conclude by discussing two modeling assumptions, namely (1) the assumption that consumers can only visit one store and (2) the assumption that firms pay an infinitesimally small setup cost for each product, and by relating our results to earlier findings in the literature on market competition with biased consumers.

1.6.1 Discussion of modeling assumptions

The impossibility of consumers to visit multiple stores may seem too restrictive at first glance. For the qualitative results and conclusions in this essay, the consequences of this assumption are in fact very mild. To see this note first that—in comparison to standard models of consumer search—the consumer in our framework has *full* information regarding her choice set when making the entry decision in stage 1: Because firms commit to perfectly observable product lines ex-ante, there is no information

to gain from visiting multiple stores. The commitment to a fixed, i.e., deterministic product line distances the fooling equilibrium also from extensively studied forms of bait-and-switch where firms limit the stock of the attraction product and then rely on positive switching cost to sell a profitable target to those customers who missed the limited bait offer (see, e.g., Lazear, 1995).

As we will now argue, the exploitation we describe in this essay does *not* rely on switching cost. The assumption that consumers visit only one store for this matter does not conceal a possible store-switching incentive on the side of consumers. The first to note is that the full information setup in our framework implies that the target must be a competitive offer in equilibrium. Because firms cannot withdraw the bait offer made to consumers ex-ante, competition is transferred into the store via the option to buy the attraction product. As in a model of direct product choice, the mark-up on the target is competed away in equilibrium. Clearly, sophisticated and rational consumers have no incentive to visit more than one store—knowing ex-ante that the choices available elsewhere do not increase their surplus. In order to study naïve consumers in a setting where switching stores is possible, one needs to define how these consumers value the product lines of other firms when preferences (unexpectedly) change due to being exposed to the local context at store k . Two possible assumptions come to mind.

The first—in our view, the more natural interpretation of context-sensitivity—is that preferences reflect a general ‘state of mind’ that applies to any options the consumer might consider when exposed to the context. In such a state of mind, options at other stores that are identical to those available at store k will be quality- or price-inflated in the exact same way as products at store k . For instance, a context might induce a quality-salient (or price-salient) state of mind, making the consumer generally willing to spend more (or less) money on a given unit of quality—regardless of where the product is located. Fixing any equilibrium we have defined in this essay, a naïve consumer would then never want to visit a second store as she does not gain a product of higher surplus elsewhere. Another possible assumption—which we find less compelling—is that the context at store k affects only the preferences over products at that store, leaving the valuation of all other products (even identical ones) unaffected. A naïve consumer might then not buy a price-salient target ($\theta_{tk}^k = P$), because she suddenly perceives the (undistorted) attraction products and targets at other stores as more valuable. If switching costs are not too high, she will want to visit more than one store. When a firm sells a *quality*-salient target ($\theta_{tk}^k = Q$),

however, the result that consumers only visit one store (where they are fooled) is robust without imposing switching costs. Because quality-salient targets are not restricted to up-selling equilibria, up-selling (with $\theta_{a^k}^k = \theta_{t^k}^k = Q$) and down-selling (with $(\theta_{a^k}^k, \theta_{t^k}^k) = (P, Q)$) predictions survive.¹¹

We have assumed that there exists an infinitesimally small cost for setting up a product. This implies that firms will not unnecessarily inflate the product line. One could argue that in reality, setup costs are either zero (in online markets) or sizable (in bricks-and-mortar markets). When setup costs are zero, all of our results continue to hold except that firms are now indifferent between setting-up profit-maximizing product lines of minimal size (which are identical with the product lines we have defined) and larger product lines that include products that have zero marginal effect on profit. Consumer choice is unaffected. We think that even without explicit setup costs, there are enough reasons for firms not to inflate the product line with options that do not affect consumer choice.¹² Of course, if setup costs are positive and sizable, fooling becomes more difficult to sustain. In this case, there will be a sufficient degree of context-sensitivity β necessary for firms to recover the additional setup cost for the un-sold attraction product (and, potentially, a decoy) with the additional fooling profit made on naïve consumers. Note that positive setup costs do not in general provide a strategic incentive to exit the market (even when profits are zero): Because the size of the product line is chosen simultaneously with other strategic variables such as qualities and prices, firms that supply the market will recover (positive but sufficiently low) setup costs with the sale price.

¹¹Of course, things become more complicated if we consider the possibility that the information of a preference change leads naïve consumers to learn something about their bias. This is an assumption that is rarely made in the literature, with Ali (2011) being a notable exception. Experiments show that people perform badly in updating beliefs about their own biases, leading us to conjecture that such effects are unlikely to make consumers fully rational. If consumers simply become more sophisticated without increasing the ability to control themselves, none of our results changes. If some consumers suddenly become rational, our results survive as long as a positive share of consumers remains naïve (see Proposition 5). A study of more involved updating procedures lies outside of the scope of this essay and is relegated to future research.

¹²Note that decoys and attraction products in our model are not unnecessary products. These products have strictly positive marginal effect on profit by enabling the fooling outcome, even in the case where no consumer purchases these products. For this reason, the minimal size of profit-maximizing product lines in the case of fooling is two (without decoys, Proposition 2) or three (with decoys, Proposition 3), respectively.

1.6.2 Related theory and findings in behavioral I.O.

There are other papers in behavioral I.O. that feature a two-phase choice procedure by which consumers first select a firm and then a product, but no study has so far considered the design of choice environments to be a source of preference distortions. Eliaz and Spiegelger (2011*b*) is related to us by the idea that ‘marketing devices’ play a role in attracting consumers to a firm. The authors study the role of zero-utility products for attracting consumers to a firm with a larger product line. At first glance, these ‘attention grabbers’ seem to be very much related to what we call the attraction product of a firm. However, there are important differences. In our model, naïve consumers mispredict their preferences and attend to the attraction product because they (falsely) expect to consume it. In Eliaz and Spiegelger (2011*b*), people have stable preferences and follow attention grabbers for reasons such as sensationalism or similarity to familiar products. As a result, Eliaz and Spiegelger (2011*b*) predict that firms use attention grabbers to attract the consumer toward products that *increase* her surplus, while we predict the opposite, namely that the use of a separate attraction product is always associated with a firm that fools consumers into buying a product of *lesser* value.¹³ Note further that a decoy, which firms in our model produce, is markedly different from the attention grabber as well. Decoys are unattractive at any stage of the decision process and therefore cannot be used to attract consumers to the firm.

Our essay compares similarly to Eliaz and Spiegelger (2011*a*) and Piccione and Spiegelger (2012). At first glance, the two papers relate to ours by the idea that ‘frames’ can influence consumer choice. At second glance, however, the mechanism of the bias and its implications are very different to salience effects in our model. Similar to attention grabbers, frames in Eliaz and Spiegelger (2011*a*) and Piccione and Spiegelger (2012) attract consumers away from status-quo products and toward products of higher value. This the reverse to how firms use salience in our model.

There are models like ours that combine a two-phase choice procedure with some form of naïve preference-distortion. These include studies of markets where firms

¹³In Eliaz and Spiegelger (2011*b*), the distortive mechanism operates over manipulating the consideration set rather than the preferences. This difference in approaches to consumer bias seems to be driving the prediction whether firms use a ‘psychology-based’ strategic variable (a.k.a. salience effects) to improve outcomes for the biased consumer (Eliaz and Spiegelger 2011*b*, for similar results see also Eliaz and Spiegelger 2011*a* and Piccione and Spiegelger 2012) or to generate possibilities to exploit them (our essay, for similar results see, e.g., Gabaix and Laibson 2006 and Heidhues and Köszegi 2010). A more in-depth analysis of this, admittedly, very interesting comparison lies however outside of the scope of this essay.

sell a bundled product that consists of a base product and a costly, unavoidable add-on (e.g., Gabaix and Laibson, 2006; Ellison, 2005), the related ‘hidden price’ literature (e.g., Heidhues, Kőszegi and Murooka, 2017), and the literature on contracting with time-inconsistent consumers (e.g., DellaVigna and Malmendier, 2004; Heidhues and Kőszegi, 2010). The studies have in common that naïve consumers miscalculate their demand (or, equivalently, the prices) at a given firm k when selecting between different suppliers. In equilibrium, profit-maximizing firms exploit this naïveté by acting as aftermarket monopolists for those consumers who experience an unexpected change to their preferences. Similar to the results in this essay, (1) competition over consumers (in the first stage) does not solve the exploitation problem, (2) the co-existence of rational and profitable-to-exploit consumers increases the problem for the exploited instead of mitigating it,¹⁴ and (3) bias-*overestimating* consumers, while also naïve, cannot be profitably exploited (see, for this particular point, Heidhues and Kőszegi, 2010).

This chapter analyzes local thinking—a widely acknowledged form of bias that has recently found formalization in theories of stimuli-driven attention such as Bordalo, Gennaioli and Shleifer (2013), Kőszegi and Szeidl (2013) and Bushong, Rabin and Schwartzstein (2016)—in markets. In our model, sellers use their product line to manipulate consumer preferences at the final point of purchase. Equilibrium marketing strategies bear strong resemblance to exploitative up- and down-selling phenomena in retail markets, with product lines that use attraction products and decoys to shift consumer attention towards profitable options. Our model predicts and explains the exploitation of naïve consumers in markets and circumstances that are not covered by the existing literature. Moreover, because salience effects that arise endogenously from the product line allows time-inconsistency to be endogenously triggered and directed by firms, we provide an extended explanation of how such biases may be formed and exploited by firms. While our model focuses on product line effects, similar incentives to design the choice environment of consumers might hold for the markets studied in other papers. In contract environments, for example, whether consumers are more or less present-biased is likely to be affected by how the terms of a contract are presented. Exploiting naïve consumers by varying the presentation of contract terms over the consumption schedule would then be very close to the salience-related fooling strategies we describe in this essay. Studying this possibility in further detail is an interesting topic for future research.

¹⁴Armstrong (2015) has recently surveyed models that make this prediction, a characteristic he calls “ripoff externalities”.

References

- Ali, S. Nageeb.** 2011. “Learning Self-Control.” *Quarterly Journal of Economics*, 126: 857–893.
- Ariely, Dan, George Loewenstein, and Drazen Prelec.** 2003. “Coherent Arbitrariness: Stable Demand Curves Without Stable Preferences.” *Quarterly Journal of Economics*, 118(1): 74–105.
- Armstrong, Mark.** 2015. “Search and Ripoff Externalities.” *Review of Industrial Organization*, 47: 273–302.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer.** 2013. “Salience and Consumer Choice.” *Journal of Political Economy*, 121(5): 803–843.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer.** 2016. “Competition for Attention.” *Review of Economic Studies*, 83: 481–513.
- Bushong, Benjamin, Matthew Rabin, and Joshua Schwartzstein.** 2016. “A Model of Relative Thinking.” mimeo (This version: March 31, 2016).
- DellaVigna, Stefano, and Ulrike Malmendier.** 2004. “Contract Design and Self-Control: Theory and Evidence.” *The Quarterly Journal of Economics*, 119(2): 353–402.
- Doyle, John R., David J. O’Connor, Gareth M. Reynolds, and Paul A. Bottomley.** 1999. “The Robustness of the Asymmetrically Dominated Effect: Buying Frames, Phantom Alternatives, and In-Store Purchases.” *Psychology and Marketing*, 16(3): 225–243.
- Eliaz, Kfir, and Ran Spiegler.** 2011*a*. “Consideration Sets and Competitive Marketing.” *Review of Economic Studies*, 78: 235–262.
- Eliaz, Kfir, and Ran Spiegler.** 2011*b*. “On the Strategic Use of Attention Grabbers.” *Theoretical Economics*, 6: 127–155.
- Ellison, Glen.** 2005. “A Model of Add-On Pricing.” *Quarterly Journal of Economics*, 120(2): 585–637.
- Ellison, Glen, and Sarah Fisher Ellison.** 2009. “Search, Obfuscation, and Price Elasticities on the Internet.” *Econometrica*, 77(2): 427–452.

- Gabaix, Xavier, and David Laibson.** 2006. “Shrouded Attributes, Consumer Myopia and Information Suppression in Competitive Markets.” *The Quarterly Journal of Economics*, 121(2): 505–540.
- Heidhues, Paul, and Botond Köszegi.** 2010. “Exploiting Naivete about Self-Control in the Credit Market.” *American Economic Review*, 100(5): 2279–2303.
- Heidhues, Paul, Botond Köszegi, and Takeshi Murooka.** 2017. “Inferior Products and Profitable Deception.” *Review of Economic Studies*, 84(1): 323–356.
- Herne, Kaisa.** 1999. “The Effects of Decoy Gambles on Individual Choice.” *Experimental Economics*, 2: 31–40.
- Huber, Joel, John W. Payne, and Christopher Puto.** 1982. “Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis.” *Journal of Consumer Research*, 9(1): 90–98.
- Jahedi, Salar.** 2011. “A Taste for Bargains.” mimeo.
- Kamenica, Emir.** 2008. “Contextual Inference in Markets: On the Informational Content of Product Lines.” *American Economic Review*, 98(5): 2127–2149.
- Köszegi, Botond, and Adam Szeidl.** 2013. “A Model of Focusing in Economic Choice.” *The Quarterly Journal of Economics*, 128(1): 53–104.
- Lazear, Edward P.** 1995. “Bait and Switch.” *Journal of Political Economy*, 103(4): 813–830.
- Mazar, Nina, Botond Köszegi, and Dan Ariely.** 2014. “True Context-Dependent Preferences? The Causes of Market-Dependent Valuations.” *Journal of Behavioral Decision Making*, 27(4): 200–208.
- Piccione, Michele, and Ran Spiegler.** 2012. “Price Competition under Limited Comparability.” *Quarterly Journal of Economics*, 127(1): 97–135.
- Simonson, Itamar.** 1989. “Choice Based on Reasons: The Case of Attraction and Compromise Effects.” *Journal of Consumer Research*, 16(2): 158–174.

Appendix to Chapter 1

We use the following method throughout all proofs to find market supply in the competitive equilibrium: First, we derive the best response of some firm k to a fixed competitor offer $\mathbf{M}^{-k} := (M^l)_{l \neq k}$ *conditional* on attracting a positive share of consumers under the assumption that the maximum price p_{\max} consumers are willing to pay is arbitrarily large, i.e., $p_{\max} \rightarrow \infty$. In general, this best response will be unique and continuous in \mathbf{M}^{-k} . Due to this characteristic, in a second step, we can find the competitive market supply by searching for the competitor offer \mathbf{M}^{-k} that equates the profits of this response to zero. At this point, firms that supply the market will sell a cost-efficient quality (q^* , q^Q , or q^P) at cost, making zero profit. When we drop the assumption $p_{\max} \rightarrow \infty$, consumers will always buy such a product if $p_{\max} \geq c(q^Q) > c(q^*) > c(q^P)$, which holds by our assumptions on the cost function (see section 1.2). The (interior) solution we define using this method is thus valid without the assumption $p_{\max} \rightarrow \infty$. Moreover, firms who do not supply the market must always choose $M^k = \emptyset$, because this is the only response that avoids any costs and yields nonnegative profits. While supplying the market at cost and choosing $M^k = \emptyset$ both yield zero profits and are thus best responses, in equilibrium, at least 2 firms must choose to supply the market. Otherwise there would exist some firm k that faced only competitors choosing $M^k = \emptyset$, making a deviation to monopoly profits possible. In general, we therefore have a range of competitive equilibria that all result in the same market supply: At least 2 firms share the market and sell at cost, while all other firms choose $M^k = \emptyset$.

Proof of Lemma 1.1 (Rational Benchmark). Let $\beta = 1$. Consumer value products according to the global surplus function, Equation (1.2). Consider some firm k and fix the competitor offer \mathbf{M}^{-k} . Let $\bar{u} \geq 0$ be the maximum surplus attainable outside of firm k (this surplus is implicitly defined by \mathbf{M}^{-k} and the outside option of no purchase). Let $p_{\max} \rightarrow \infty$ and consider the best response *conditional* on attracting a positive share of consumers. Fix some quality $q_j \geq q_{\min}$. The firm can sell q_j to all consumers at price $p_j = \lim_{\delta \rightarrow 0} (q_j - \bar{u} - \delta) = q_j - \bar{u} \Leftrightarrow u_j = \bar{u}$, where $\delta > 0$ is the smallest monetary unit. At this price, the firm offers just enough surplus to let consumers marginally improve over the highest surplus available elsewhere, thereby winning all consumers. For given quality q_j , no other price can achieve higher profits: A higher price implies the loss of all consumers, a lower price cannot attract more. This price implies profit $\pi^k = q_j - \bar{u} - c(q_j)$ and thus, the profit-maximizing quality

to sell is $q^* := \arg \max[q - c(q)]$, or $c'(q^*) = 1$. Note that $q^* > q_{\min}$ by assumption, making this interior solution valid. Offering additional products is costly and cannot increase profits. It follows: Conditional on attracting a positive share of consumers, the unique best response is the product line $M^k = ((q^*, q^* - \bar{u}))$. Note that the best response so defined is unique and continuous in \bar{u} . Market supply in the competitive equilibrium can thus be found by searching for \bar{u} where this response yields zero profits. This unique point exists at $\bar{u} = q^* - c(q^*)$, implying marginal cost pricing, $p_j = c(q^*)$ and the product line $M^* = ((q^*, c(q^*)))$. This solution is valid by our model assumption $p_{\max} > c(q^*)$, such that we can drop the assumption $p_{\max} \rightarrow \infty$.

Given that some firm offers $M^* = ((q^*, c(q^*)))$, other firms face $\bar{u} = q^* - c(q^*)$. There are two best responses: (1) Sell $M^* = ((q^*, c(q^*)))$ as well, which yields zero profits, (2) Offer nothing, $M^k = \emptyset$, which is the only response avoiding all costs and also yields zero profits. In any equilibrium, at least 2 firms must offer the product line M^* : If no firm offered M^* , then any firm would face an outside option $\bar{u} = 0 < q^* - c(q^*)$ and there would exist a deviation incentive to monopoly profits. If only one firm offered M^* , then, similarly, this firm could earn monopoly profits by deviating. We thus have a range of competitive equilibria that all result in the same market supply: At least 2 firms share the market and offer M^* , while all other firms choose $M^k = \emptyset$. \square

Proof of Proposition 1.1 (Sophisticated Populations). Let $\beta > 1$. Assume that $\tilde{\beta} = \beta$ for all consumers. All consumers have correct expectations about their in-store preferences. They enter store k if and only if the purchase at store k yields higher *global* surplus (Equation(1.2)) than the outside option and the expected purchase elsewhere. Let $\bar{u} \geq 0$ be the maximum global surplus attainable outside of firm k . Assume $p_{\max} \rightarrow \infty$. As in the rational benchmark, the firm can sell quality $q_j \geq q_{\min}$ to all consumers at price $p_j = \lim_{\delta \rightarrow 0}(q_j - \bar{u} - \delta) = q_j - \bar{u} \Leftrightarrow u_j = \bar{u}$, where $\delta > 0$ is the smallest monetary unit. It follows that the profit maximizing quality to sell is $q_j = q^*$. Conditional on attracting a positive share of consumers, the unique best response is the product line $M^k = ((q^*, q^* - \bar{u}))$. This is identical to the unique best response if consumers are rational (Lemma 1). Market supply in the equilibrium is thus identical to the rational benchmark. \square

Proof of Lemma 1.2 (Profitable Fooling). Assume $\beta > 1$ and consider a naïve local thinker with $\tilde{\beta} < \beta$. Consider a profit-maximizing firm that offers a single product of quality $q_j > q_{\min}$, which it sells at price $p_j < p_{\max}$ to the consumer. Profit maximization (conditional on offering a single product j) implies that

$p_j = \lim_{\delta \rightarrow 0}(q_j - \bar{u} - \delta) = q_j - \bar{u}$, where $\bar{u} \geq 0$ is the maximum global surplus that the consumer expects to attain elsewhere.

Fix \bar{u} , q_j and p_j and assume that, instead, the firm would offer a product line with two products, a and t , $M^k = ((q_a, p_a), (q_t, p_t))$. Assume that $(\theta_a^k, \theta_t^k) = (Q, Q)$. The consumer is attracted by product a , but purchases product t (the consumer is fooled) if and only if (IC) $\beta q_t - p_t \geq \beta q_a - p_a$ and (PCC) $\tilde{\beta} q_t - p_t \leq \tilde{\beta} q_a - p_a$, with at least one of the inequalities being strict. By $\tilde{\beta} < \beta$, this requires $q_t > q_a$ and $p_t > p_a$ (the firm up-sells). Conditional on selling product t , an upper bound on the price of product t is given by $\beta q_t - p_t = \beta q_a - p_a \Leftrightarrow p_t = \beta(q_t - q_a) + p_a$. Conditional on attracting the consumer, an upper bound on price p_a is given by $u_a = \bar{u} \Leftrightarrow p_a = q_a - \bar{u}$. At these prices, (IC) holds with strict equality and (PCC) with strict inequality for any $\tilde{\beta} < \beta$. Fix these prices and choose $q_t = q_j$. Then $p_t > p_j$ if and only if $q_a < q_j$. Note that $p_t = \beta(q_j - q_a) + q_a - \bar{u}$ is continuous and strictly increasing in $(q_j - q_a)$, with $\lim_{q_a \rightarrow q_j} p_t = p_j$. From $q_j > q_{\min}$ and $p_j < p_{\max}$ it then follows that there exist a range of $q_a \in [q_{\min}, q_j)$ for which it holds that $p_j < p_t < p_{\max}$: By an adequate choice of q_a , the firm can fool the consumer and sell quality q_j at a strictly higher price $p_t > p_j$ (and thus, profit) than by offering product line $M^k = ((q_j, p_j))$.

The proofs for $(\theta_a^k, \theta_t^k) \in \{(P, P), (P, N), (N, Q), (P, Q)\}$ are analogous to the case of $(\theta_a^k, \theta_t^k) = (Q, Q)$: Fixing price p_t such that (IC) binds ($u_t^k(\beta) = u_a^k(\beta)$) and price p_a to $p_a = q_a - \bar{u}$, there exists range of qualities $q_a \geq q_{\min}$ which allow the firm to fool the consumer and sell quality $q_t = q_j$ at price $p_t > p_j$. Note that if $(\theta_a^k, \theta_t^k) = (P, P)$, (IC) $q_t - \beta p_t \geq q_a - \beta p_a$ and (PCC) $q_t - \tilde{\beta} p_t \leq q_a - \tilde{\beta} p_a$. Fooling then implies that $q_t < q_a$ and $p_t < p_a$, i.e., that the firm down-sells. If the distortion is asymmetric, $(\theta_a^k, \theta_t^k) \in \{(P, N), (N, Q), (P, Q)\}$, (IC) and (PCC) do *not* constrain qualities q_a, q_t and prices p_a, p_t to a particular order. More precisely, the interval of qualities q_a that generate a fooling outcome is then bound below by some quality $\underline{q} < q_t$ (allowing for q_a that generate an up-sell) and above by some quality $\bar{q} > q_t$ (allowing for q_a that generate a down-sell).

It remains to be shown that $(\theta_a^k, \theta_t^k) \in \{(Q, Q), (P, P), (P, N), (N, Q), (P, Q)\}$ are the only pairs of distortions that can generate a profitable fooling outcome. The result is immediate if we try proving the profitability of pair $(\theta_a^k, \theta_t^k) \in \{(Q, N), (N, P), (Q, P)\}$ analogous to the case of $(\theta_a^k, \theta_t^k) = (Q, Q)$. We show this exemplarily for $(\theta_a^k, \theta_t^k) = (N, P)$. If $(\theta_a^k, \theta_t^k) = (N, P)$, fooling requires that (IC) $q_t - \beta p_t \geq q_a - p_a$ and (PCC) $q_t - \tilde{\beta} p_t \leq q_a - p_a$, with at least one of the inequalities being strict. But this requires that $p_t < 0$, which obviously cannot be

profitable. Similar results obtain for $(\theta_a^k, \theta_t^k) \in \{(Q, N), (Q, P)\}$. \square

Proof of Proposition 1.2 (Fooling with an unspecified salience technology.) We derive the equilibrium from the best response of a given firm k to a generic market situation. For ease of notation, we drop the superscript k from products t^k and a^k . We begin the proof by considering a perfectly homogeneous and naïve consumer population with unique type $\tilde{\beta}^0 < \beta$. Consider a generic firm k . Fix the competitor offer \mathbf{M}^{-k} and let $\bar{u} = \bar{u}(\tilde{\beta}^0) \geq 0$ be type $\tilde{\beta}^0$'s *expected* maximum surplus attainable outside of firm k . Assume (for now) that $p_{\max} \rightarrow \infty$.

We derive the best response *conditional* on attracting a positive share of consumers. Lemma 1.2 has established the profitability of fooling strategies over the entire range of possible naïveté $\tilde{\beta}$, interior quality $q > q_{\min}$ and interior price $p < p_{\min}$. It follows that if a best response exist, it must involve fooling. Assume that the firm fools, selling product t , but attracting the consumer with product $a \neq t$. The maximum price the firm can sell target quality q_t obtains from setting $u_t^k(\beta) = u_a^k(\beta)$ ((IC) binds) while setting the quality and price of the attraction product such that $u_a = \bar{u}$ (the participation constraint binds). At this price, (PCC) is slack for any $\tilde{\beta} < \beta$, implying that the consumer is fooled. To achieve this price, offering two products is necessary and sufficient. If the firm can choose distortions θ_j^k independently from the product line, holding more than 2 products is unnecessary yet costly and can thus not be part of the best response. So $M^k = ((q_a, p_a), (q_t, p_t))$. When firms are restricted to symmetric distortions, the firm chooses either $(\theta_a^k, \theta_t^k) = (Q, Q)$ or $(\theta_a^k, \theta_t^k) = (P, P)$. If firms are able to choose asymmetric distortions, it is easy to see that the unique profit maximizing choice is $(\theta_a^k, \theta_t^k) = (P, Q)$: Such a distortion maximizes the wedge between the utility difference $u_t - u_a$ (outside the store) and the utility difference $u_t^k(\beta) - u_a^k(\beta)$ (inside the store). For given target quality q_t , the distortion $(\theta_a^k, \theta_t^k) = (P, Q)$ therefore maximizes the selling price p_t in a fooling situation.

- Best response if $(\theta_a^k, \theta_t^k) = (Q, Q)$. The profit-maximizing price for the target is $p_t = \beta(q_t - q_a) + p_a$ ((IC) binds) under the condition that $q_a - p_a = \bar{u}$ (the participation constraint binds). With the quality of the target being inflated at the store, the cost-efficient quality to sell is $q_t = q^Q := \arg \max_q [\beta q - c(q)] \Leftrightarrow c'(q^Q) = \beta$. This interior solution is valid by assumption $q^Q > q_{\min}$. We are left with the choice of the attraction product (q_a, p_a) . There are 2 opposing options: Minimizing q_a and maximizing p_a . The profit-maximizing choice is to minimize q_a : Because quality q_a is inflated at

the store, the positive effect on profits of decreasing quality q_a is larger than the positive effect of increasing price p_a . The unique profit-maximizing choice is therefore to choose $q_a = q_{\min}$, which implies $p_a = q_{\min} - \bar{u}$. Note that $q_a < q_t$, $p_a < p_t$ and $u_t < u_a$. The best response is characterized by:

$$\begin{aligned} (\theta_a^k, \theta_t^k) &= (Q, Q), (q_t, p_t) = (q^Q, \beta q^Q - (\beta - 1)q_{\min} - \bar{u}), \\ (q_a, p_a) &= (q_{\min}, q_{\min} - \bar{u}) \end{aligned} \quad (\text{Q})$$

- Best response if $(\theta_a^k, \theta_t^k) = (P, P)$. The profit-maximizing price for the target is $p_t = p_a - \frac{1}{\beta}(q_a - q_t)$ ((IC) binds) under the condition that $q_a - p_a = \bar{u}$ (the participation constraint binds). With the price of the target being inflated at the store, the cost-efficient quality to sell is $q_t = q^P := \arg \max_q [q - \beta c(q)] \Leftrightarrow c'(q^P) = \frac{1}{\beta}$. This interior solution is valid by assumption $q^P \geq q_{\min}$. We are left with the choice of the attraction product (q_a, p_a) . There are 2 opposing options: Minimizing q_a and maximizing p_a . The profit-maximizing choice now is to maximize p_a : Because price p_a is inflated at the store, the positive effect on profits of increasing price p_a is larger than the positive effect of decreasing quality q_a . The unique profit-maximizing choice is therefore to choose $p_a = p_{\max}$, which implies $q_a = p_{\max} + \bar{u}$. Note that $q_a > q_t$, $p_a > p_t$ and $u_t < u_a$. The best response is characterized by:

$$\begin{aligned} (\theta_a^k, \theta_t^k) &= (P, P), (q_t, p_t) = (q^P, p_{\max} - \frac{1}{\beta}(p_{\max} + \bar{u} - q^P)), \\ (q_a, p_a) &= (p_{\max} + \bar{u}, p_{\max}) \end{aligned} \quad (\text{P})$$

- Best response if $(\theta_a^k, \theta_t^k) = (P, Q)$. Then the profit-maximizing price for the target is $p_t = \beta q_t - q_a + \beta p_a$ ((IC) binds) under the condition that $q_a - p_a = \bar{u}$ (the participation constraint binds). With the quality of the target being inflated at the store, the cost-efficient quality to sell is $q_t = q^Q := \arg \max_q [\beta q - c(q)] \Leftrightarrow c'(q^Q) = \beta$. This interior solution is valid by assumption $q^Q > q_{\min}$. We are left with the choice of the attraction product (q_a, p_a) . There are 2 opposing options: Minimizing q_a and maximizing p_a . The profit-maximizing choice now is to maximize p_a : Because price p_a is inflated at the store, the positive effect on profits of increasing price p_a is larger than the positive effect of decreasing quality q_a . The unique profit-maximizing choice is therefore to choose $p_a = p_{\max}$, which implies $q_a = p_{\max} + \bar{u}$. Note that $u_t < u_a$.

The best response is characterized by:

$$\begin{aligned} (\theta_a^k, \theta_t^k) &= (P, Q), (q_t, p_t) = (q^Q, \beta q^Q + (\beta - 1)p_{\max} - \bar{u}), \\ (q_a, p_a) &= (p_{\max} + \bar{u}, p_{\max}) \end{aligned} \quad (\text{PQ})$$

Note that the best response in all three cases is independent of the degree of naïveté of type $\tilde{\beta}^0 < \beta$: Due to the optimality condition $u_t^k(\beta) = u_a^k(\beta)$ ((IC) binds), any consumer with belief $\tilde{\beta} < \beta$ (falsely) believes to purchase product a with certainty. The best response does *not* generate heterogeneous expectations among a population that contains heterogenous degrees of naïveté. If firms play mutual best responses, any heterogeneity in types $\tilde{\beta}$ is therefore rendered unimportant for market supply: Uniqueness of the best response implies that firms generating positive demand must choose according to it; otherwise, there would exist a strict deviation incentive. This response does not generate heterogeneous expectations. Firms not generating positive demand, on the other hand, choose $M^k = \emptyset$ to avoid positive costs and thus negative profits. These firms do not generate heterogeneous expectations either. It follows that in any equilibrium, $\bar{u}(\tilde{\beta}) = \bar{u} \forall \tilde{\beta} < \beta$: the outside option is a unique value. Equilibrium candidates are independent of the distribution of naïveté and can be derived by finding the (unique) value for \bar{u} that equates best response profits to zero. This yields the following equilibrium candidates:

$$\begin{aligned} (\theta_a, \theta_t) &= (Q, Q), (q_t, p_t) = (q^Q, c(q^Q)), (q_a, p_a) = (q_{\min}, c(q^Q) - \beta(q^Q - q_{\min})) \quad (\text{Q}^*) \\ (\theta_a, \theta_t) &= (P, P), (q_t, p_t) = (q^P, c(q^P)), (q_a, p_a) = (q^P + [p_{\max} - c(q^P)], p_{\max}) \quad (\text{P}^*) \\ (\theta_a, \theta_t) &= (P, Q), (q_t, p_t) = (q^Q, c(q^Q)), (q_a, p_a) = (\beta(q^Q + p_{\max}) - c(q^Q), p_{\max}) \end{aligned} \quad (\text{PQ}^*)$$

In equilibrium, at least two firms must provide a non-empty product line according to the respective candidate. These firms share the market. All other firms choose $M^k = \emptyset$. Note that the characterizations are valid for any $p_{\max} \geq c(q^Q) > c(q^P)$ as assumed in the model section of this chapter. We can drop the assumption that $p_{\max} \rightarrow \infty$.

We are ready to characterize the equilibrium. If firms can choose asymmetric distortions, the unique best response involves choosing $(\theta_a, \theta_t) = (P, Q)$ and thus, equilibrium product supply is uniquely characterized by (PQ^{*}). If firms are restricted to choosing symmetric distortions, both up-selling equilibria (Q^{*}) and down-selling equilibria (P^{*}) can emerge. Fix a quality-salient equilibrium according to (Q^{*}) and

consider some firm k . There exists at least one firm $l \neq k$ with a product line $M^l = ((q_a, p_a), (q_t, p_t))$ and product characteristics defined according to (Q*). This firm provides expected surplus $\bar{u} = u_a = q_{\min} - c(q^Q) + \beta(q^Q - q_{\min})$ to the naïve consumer population. If (Q*) indeed defines an equilibrium, firm k either provides a product line that is identical to the product line of firm l or an empty product line. The only profitable deviation that might exist is a deviation towards a price-salient store with $(\theta_a^k, \theta_t^k) = (P, P)$. The most profitable deviation is given by the best response we have derived above: The firm offers a product-line with two products, $M^k = ((q_a, p_a), (q_t, p_t))$ satisfying $(q_t, p_t) = (q^P, p_{\max} - \frac{1}{\beta}(p_{\max} + \bar{u} - q^P))$ and $(q_a, p_a) = (p_{\max} + \bar{u}, p_{\max})$, with $\bar{u} = q_{\min} - c(q^Q) + \beta(q^Q - q_{\min})$. This deviation is strictly profitable if and only if it yields $q_t - p_t > 0$. Rearranging, this is the case if and only if $\nu^{(Q,Q)} < \nu^{(P,P)}$, where

$$\begin{aligned}\nu^{(Q,Q)} &:= (q^Q - c(q^Q)) + (\beta - 1)(q^Q - q_{\min}), \text{ and} \\ \nu^{(P,P)} &:= (q^P - c(q^P)) + (\beta - 1)(p_{\max} - c(q^P)).\end{aligned}$$

Analogously, in a price-salient equilibrium characterized by (P*), firms have a deviation incentive to a quality-salient store $(\theta_a^k, \theta_t^k) = (Q, Q)$ if and only if $\nu^{(Q,Q)} > \nu^{(P,P)}$. We conclude: If $\nu^{(Q,Q)} > \nu^{(P,P)}$, equilibrium product supply follows (Q*). If $\nu^{(Q,Q)} < \nu^{(P,P)}$, equilibrium product supply follows (P*) In the knife-edge case of $\nu^{(Q,Q)} = \nu^{(P,P)}$, product supply can either follow (Q*) or (P*). \square

Proof of Proposition 1.3 (Fooling with Salience, Focusing, or Relative Thinking).

Assume that θ_j^k follows Assumption BGS, KS or BRS. We show that using a product-line with three products $M^k = ((q_a, p_a), (q_t, p_t), (q_d, p_d))$ is necessary and sufficient to enable a fooling strategy identical to the case when firms choose distortions θ_j^k directly. In particular, we show that product d is necessary and sufficient to let firms choose (q_a, p_a) , (q_t, p_t) and (θ_a^k, θ_t^k) according to the best response we have defined in the proof of Proposition 1.2. Assumptions KS and BRS restrict firms to symmetric distortions. The best response then involves using product d to construct either a quality-salient store, characterization (Q), or a price-salient store, characterization (P). Assumption BGS allows firms to choose asymmetric distortions. In this case, firms best respond with a product line where product d is used to construct a fooling regime according to characterization (PQ).

1. *Assumption KS.* Consider the best response characterized by (Q) or (P), proof of Proposition 1.2.

- (A decoy is necessary.) Assume $M^k = ((q_a, p_a), (q_t, p_t))$. We show that under Assumption KS, fooling is generally impossible with a product-line of two products. If $(\theta_a^k, \theta_t^k) = (Q, Q)$, fooling requires $q_t > q_a, p_t > p_a$, (IC) $\beta(q_t - q_a) \geq p_t - p_a$ and (PCC) $\tilde{\beta}(q_t - q_a) \leq p_t - p_a$. By $\tilde{\beta} < \beta$, (IC) and (PCC) together imply $q_t - q_a \leq p_t - p_a$. However, Assumption KS requires $q_t - q_a > p_t - p_a$ for $(\theta_a^k, \theta_t^k) = (Q, Q)$, a contradiction. If $(\theta_a^k, \theta_t^k) = (P, P)$, fooling requires $q_t < q_a, p_t < p_a$, (IC) $\beta(p_a - p_t) \geq q_a - p_t$ and (PCC) $\tilde{\beta}(p_a - p_t) \leq q_a - p_t$. By $\tilde{\beta} < \beta$, (IC) and (PCC) together imply $p_t - p_a \leq q_t - q_a$. However, Assumption KS requires $p_a - p_t > q_a - q_t$ for $(\theta_a^k, \theta_t^k) = (P, P)$, a contradiction.¹⁵
- (One decoy is sufficient.) Assume $M^k = ((q_a, p_a), (q_t, p_t), (q_d, p_d))$ and consider a best response according to (Q). Choose, for example, $p_d = p_t$ and $q_d < q_t - (p_t - p_a)$. Then $(\max_{j \in J^k} q_j - \min_{j \in J^k} q_j) = q_t - q_d$, $(\max_{j \in J^k} p_j - \min_{j \in J^k} p_j) = p_t - p_a$ and $(\max_{j \in J^k} q_j - \min_{j \in J^k} q_j) > (\max_{j \in J^k} p_j - \min_{j \in J^k} p_j)$. By Assumption KS, $(\theta_a^k, \theta_t^k) = (Q, Q)$. Note that $u_a^k(\tilde{\beta}) < u_a^k(\beta)$ and $u_d^k(\tilde{\beta}) < u_t^k(\beta)$: Adding product d allows the construction of $(\theta_a^k, \theta_t^k) = (Q, Q)$ without inducing a violation of (PCC) or (IC). The construction is analogous with a best response according to (P): To implement $(\theta_a^k, \theta_t^k) = (P, P)$, choose, for example, $q_d = q_t$ and $p_d > p_t + (q_a - q_t)$.

2. *Assumption BRS.* Consider the best response characterized by (Q) or (P), proof of Proposition 1.2.

- (A decoy is necessary.) The best response involves fooling. We show that under Assumption BRS, fooling is generally impossible with a product-line of two products. Assume $M^k = ((q_a, p_a), (q_t, p_t))$. If $(\theta_a^k, \theta_t^k) = (Q, Q)$, fooling requires $q_t > q_a, p_t > p_a$, (IC) $\beta(q_t - q_a) \geq p_t - p_a$ and (PCC) $\tilde{\beta}(q_t - q_a) \leq p_t - p_a$. However, Assumption BRS requires $\beta(q_t - q_a) < p_t - p_a$ for $(\theta_a^k, \theta_t^k) = (Q, Q)$, a contradiction of (IC). If $(\theta_a^k, \theta_t^k) = (P, P)$, fooling requires $q_t < q_a, p_t < p_a$, (IC) $\beta(p_a - p_t) \geq q_a - p_t$ and (PCC) $\tilde{\beta}(p_a - p_t) \leq q_a - p_t$. However, Assumption BRS requires $\beta(p_a - p_t) < q_a - p_t$ for $(\theta_a^k, \theta_t^k) = (P, P)$, a contradiction of (IC).¹⁶

¹⁵This result follows from a general characteristic of the Focusing framework: If there are just two options, focusing weights favor the option that would also be chosen by a rational consumer (a simple corollary of Proposition 3 (“balanced tradeoffs”) in Kőszegi and Szeidl, 2013). In our framework this implies that in a product line with just two products, if $u_a \geq u_t$, then $u_a^k(\beta) > u_t^k(\beta)$, rendering fooling impossible.

¹⁶This result follows immediately from Norming Assumption N2 in Bushong, Rabin and

- (One decoy is sufficient.) Assume $M^k = ((q_a, p_a), (q_t, p_t), (q_d, p_d))$ and consider a best response according to (Q). Choose, for example, $q_d = q_t$ and $p_d > p_a + \beta(q_t - q_a)$. Then $(\max_{j \in J^k} q_j - \min_{j \in J^k} q_j) = q_t - q_a$, $(\max_{j \in J^k} p_j - \min_{j \in J^k} p_j) = p_d - p_a$ and $(\max_{j \in J^k} p_j - \min_{j \in J^k} p_j) > \beta(\max_{j \in J^k} q_j - \min_{j \in J^k} q_j)$. By Assumption BRS, $(\theta_a^k, \theta_t^k) = (Q, Q)$. Note that $u_d^k(\tilde{\beta}) < u_a^k(\tilde{\beta})$ and $u_d^k(\beta) < u_t^k(\beta)$: Adding product d allows the construction of $(\theta_a^k, \theta_t^k) = (Q, Q)$ without inducing a violation of (PCC) or (IC). The construction is analogous with a best response according to (P): To implement $(\theta_a^k, \theta_t^k) = (P, P)$, choose, for example, $p_d = p_t$ and $q_d < q_a - \beta(p_a - p_t)$.

3. *Assumption BGS.* Consider the best response characterized by (PQ), proof of Proposition 1.2.

- (A decoy is necessary.) Assume $M^k = ((q_a, p_a), (q_t, p_t))$. Best response (PQ) implies $q_a > q_t$ and $p_a > p_t$. Thus, none of the two products is dominated. The reference quality is given by $\bar{q}^k = \frac{(q_a + q_t)}{2}$ and the reference price is given by $\bar{p}^k = \frac{(p_a + p_t)}{2}$. Because $(q_j - \bar{q}^k)(p_j - \bar{p}^k) > 0$ for $j = a, t$, we can exploit Proposition 1 in Bordalo, Gennaioli and Shleifer (2013): The “advantageous” attribute of product j —higher quality *or* lower price relative to the reference—is overweighted if and only if $\frac{q_j}{p_j} > \frac{\bar{q}^k}{\bar{p}^k}$. Also, if and only if $\frac{q_j}{p_j} < \frac{\bar{q}^k}{\bar{p}^k}$, then the “disadvantageous” attribute of product j is overweighted, while if and only if $\frac{q_j}{p_j} = \frac{\bar{q}^k}{\bar{p}^k}$, consumers weigh both attributes equally. Assume towards a contradiction that the firm can construct $(\theta_a, \theta_t) = (P, Q)$. For t being quality-salient, by $q_t < \bar{q}^k$ and Proposition 1 in BGS,

$$\frac{q_t}{p_t} < \frac{\bar{q}^k}{\bar{p}^k} \Leftrightarrow \frac{q_t}{p_t} < \frac{q_a}{p_a}.$$

But for a being price-salient, by $q_a > \bar{q}^k$ and Proposition 1 in BGS,

$$\frac{q_a}{p_a} < \frac{\bar{q}^k}{\bar{p}^k} \Leftrightarrow \frac{q_t}{p_t} > \frac{q_a}{p_a},$$

a contradiction.

Schwartzstein (2016), which implies that in choice sets with just two options (that differ on two dimensions), “relative thinkers” behave as if maximizing an unweighted utility function: See the discussion on page 7 in Bushong, Rabin and Schwartzstein (2016).

- (One decoy is sufficient.) Assume $M^k = ((q_a, p_a), (q_t, p_t), (q_d, p_d))$. Best response (PQ) implies $q_a > q_t > q_{\min} > 0$ and $p_a > p_t > 0$.

Assume that $\frac{q_t}{p_t} > \frac{q_a}{p_a}$. We construct a reference point using product d that satisfies the following properties: (1) $\bar{p}^k = p_t$, (2) $\bar{q}^k < q_t$ and (3) $\frac{q_a}{p_a} < \frac{\bar{q}^k}{\bar{p}^k} < \frac{q_t}{p_t}$. The construction is illustrated in Figure 1.3 (left panel). With such a reference point,

1. Product t is quality-salient: By $\bar{p}^k = p_t$, the salience of p_t is $\sigma(p_t, p_t)$. By homogeneity of degree zero, $\sigma(\alpha p_t, \alpha p_t) = \sigma(p_t, p_t)$ for any $\alpha > 0$. Let $\alpha = \frac{q_t}{p_t} > 0$, then $\sigma(p_t, p_t) = \sigma(q_t, q_t)$. By ordering, $\sigma(q_t, q_t) < \sigma(q_t, \bar{q}^k)$ because $\bar{q}^k < q_t$. Thus, $\sigma(q_t, \bar{q}^k) > \sigma(p_t, \bar{p}^k)$: product t is quality-salient.
2. Product a is price-salient: By $\bar{q}^k < q_t < q_a$ and $\bar{p}^k = p_t < p_a$, $(q_a - \bar{q}^k)(p_a - \bar{p}^k) > 0$, and product a neither dominates nor is dominated by the reference good. Thus, Proposition 1 in BGS applies. Because $q_a > \bar{q}^k$, by $\frac{\bar{q}^k}{\bar{p}^k} > \frac{q_a}{p_a}$, product a is price-salient.

To satisfy property (1), choose $p_d = 2p_t - p_a$, which implies $p_d < p_t$. To satisfy property (2) and (3), choose $q_d < 2q_t - q_a$, which implies $q_d < q_t$. It remains to be shown that the decoy d does not violate fooling conditions. Note that $q_d - p_d < 2q_t - q_a - (2p_t - p_a) \Leftrightarrow u_d < 2u_t - u_a$. Because $u_t < u_a$ by the specifications of a and t , this implies that $u_d < u_t < u_a$. We first show that (IC) is not violated: Because t is quality-salient, $u_t^k(\beta) = \beta q_t - p_t > u_t$. But then, if (i) $\theta_d^k = N$, $u_t^k(\beta) > u_d^k(\beta)$ follows from $u_t^k(\beta) > u_t > u_d = u_d^k(\beta)$, if (ii) $\theta_d^k = Q$, $u_t^k(\beta) > u_d^k(\beta)$ follows from $q_d < q_t$, $p_d < p_t$ and $u_t > u_d$, if (iii) $\theta_d^k = P$, then $u_t^k(\beta) > u_d^k(\beta)$ if and only if $u_a^k(\beta) > u_d^k(\beta) \Leftrightarrow q_a - q_d > \beta(p_a - p_d)$ by $u_t^k(\beta) = u_a^k(\beta)$. To prove that $q_a - q_d > \beta(p_a - p_d)$, note that $q_a - q_d > q_a - (2q_t - q_a) = 2(q_t - q_a)$ by $q_d < 2q_t - q_a$ and $p_a - p_d = p_a - (2p_t - p_a)$ by $p_d = 2p_t - p_a$. Thus $q_a - q_d > \beta(p_a - p_d)$ if $2(q_a - q_t) > 2\beta(p_a - p_t) \Leftrightarrow (q_a - q_t) > \beta(p_a - p_t)$. But the latter inequality is true by $u_t^k(\beta) = u_a^k(\beta) \Leftrightarrow q_a - \beta q_t = \beta p_a - p_t$. Thus, $u_t^k(\beta) > u_d^k(\beta)$. Finally, we have to show that (PCC) is not violated, i.e., that $u_a^k(\tilde{\beta}) > u_d^k(\tilde{\beta})$. To see that this is true note that we have shown that $u_a > u_t > u_d$ and $u_a^k(\beta) = u_t^k(\beta) > u_d^k(\beta)$. Because $u_a^k(\tilde{\beta})$ is between $u_a^k(\beta)$ and u_a and $u_d^k(\tilde{\beta})$ is between $u_d^k(\beta)$ and u_d (both by $\tilde{\beta} < \beta$) it follows that $u_a^k(\tilde{\beta}) > u_d^k(\tilde{\beta})$.

Assume that $\frac{q_t}{p_t} < \frac{q_a}{p_a}$. We construct a reference point using one additional product d that satisfies the following properties: (1) $\bar{q}^k = q_a$, (2) $\bar{p}^k > p_a$ and

(3) $\frac{q_a}{p_a} > \frac{\bar{q}^k}{\bar{p}^k} > \frac{q_t}{p_t}$. The construction is illustrated in Figure 1.3 (right panel). With such a reference point,

1. Product t is quality-salient: By $\bar{q}^k > q_t$ and $\bar{p}^k > p_t$, $(q_t - \bar{q}^k)(p_t - \bar{p}^k) > 0$, and product t neither dominates nor is dominated by the reference good. Thus, Proposition 1 in BGS applies. Because $q_t < \bar{q}^k$, by $\frac{\bar{q}^k}{\bar{p}^k} > \frac{q_t}{p_t}$, product t is quality-salient.
2. Product a is price-salient: By $\bar{p}^k = p_a$, the salience of q_a is $\sigma(q_a, q_a)$. By homogeneity of degree zero, $\sigma(\alpha q_a, \alpha q_a) = \sigma(q_a, q_a)$ for any $\alpha > 0$. Let $\alpha = \frac{p_a}{q_a} > 0$, then $\sigma(q_a, q_a) = \sigma(p_a, p_a)$. By ordering, $\sigma(p_a, p_a) < \sigma(p_a, \bar{p}^k)$ because $\bar{p}^k > p_a$. Thus, $\sigma(q_a, \bar{q}^k) < \sigma(p_a, \bar{p}^k)$: product a is price-salient.

To satisfy property (1) choose $q_d = 2q_a - q_t > q_a$. To satisfy property (2) and (3), choose $p_d > 2p_a - p_t$. It remains to be shown that the decoy d does not violate fooling conditions. But note that $p_d > p_a = b$: The decoy has a price above the maximum willingness to pay and thus, will never be chosen (and can therefore not violate fooling conditions).

□

Proof of Proposition 1.4 (Co-Existence of Sophisticated and Naïve Agents). Let $\beta > 1$. Fix market supply according to the Proposition. There exist two types of stores with strictly positive demand, k^F and k^* . Type k^F is a fooling firm that supplies products according to the equilibrium defined in Proposition 1.2 and k^* is a non-fooling firm that supplies products according to the rational benchmark Lemma 1.1. There exist at least two firms of each type. All other firms choose $M^k = \emptyset$. All firms make zero profits. Note that conditional on purchasing at type k^* , all consumers expect to purchase q^* at price $p^* = c(q^*)$, yielding utility $u^* = q^* - c(q^*)$. At the same time, conditional on purchasing at type k^L , all sophisticated consumers (correctly) expect to purchase the target (yielding utility $u_t = q_t - c(q_t)$), while all naïves (falsely) expect to purchase the attraction product (yielding utility $u_a = q_a - p_a > u_t$). We prove that a competitive equilibrium with this market supply exists and that it defines the unique competitive market supply.

(Existence.) Assume that we have an equilibrium. Firms of type k^L fool and sell quality $q_t \neq q^*$ at $p_t = c(q_t)$ to the naïves, while firms of type k^H do not fool and sell q^* at $p^* = c(q^*)$ to the sophisticated consumers. We have to check whether consumers or firms want to deviate. Consumers do not want to deviate: By the

strict convexity of the cost function, $u_a > u^* > u_t$. The first inequality prevents naïves to purchase at k^* , the second inequality prevents sophisticated consumers to purchase at k^F . Firms of either type also do not have an incentive to deviate. By Proposition 1.2, no firm can find a more profitable strategy when serving naïves if there are at least two firms of type k^F in the market. By Proposition 1.1, no firm can find a more profitable strategy when serving sophisticated agents if there exist at least two firms of type k^* .

(Uniqueness.) The proofs of Propositions 1.1 and 1.2, respectively, show that unless there exist at least two firms supplying products according to Proposition 1.1 as well as at least two firms supplying products according to Proposition 1.2, there exists a deviation incentive to a strategy with strictly positive profits. In particular, by the uniqueness and continuity of the best response conditional on attracting only sophisticated consumers (Proposition 1.1), there must exist at least two firms supplying a product with expected surplus $\bar{u}^* \geq u^* = q^* - c(q^*)$ to consumers of type $\tilde{\beta} \geq \beta$. Otherwise, at least one firm could attract the entire population of types $\tilde{\beta} \geq \beta$ at strictly positive profit. Similarly, there must exist at least two firms supplying a product with expected surplus $\bar{u}^F \geq u_a = q_a - p_a$ to consumers of type $\tilde{\beta} < \beta$, where q_a and p_a are defined by the equilibrium characterized in Proposition 1.2. Otherwise, at least one firm could attract the entire population of types $\tilde{\beta} < \beta$ at strictly positive profit. By the strict difference of u_a and u^* (in particular, $u_a > u^*$), a single firm cannot satisfy both of these conditions at the same time (attracting both groups of consumers with positive probability), even if it would play a mixed strategy: Such a firm would either have to make negative profits in expectation (to attract both groups without generating a deviation incentive for other firms) or generate an offer that (for at least one of the two groups of consumers) could be profitably undercut by other firms. It follows that at least two firms satisfying the respective condition must exist for each group *separately*. Because each firm only serves one group of consumers, the only possibility to satisfy the respective condition without making negative profit is for each firm to choose market supply according to Propositions 1.1 and 1.2, respectively. It follows that any competitive equilibrium must have the characteristics listed in the Proposition. \square

Proof of Proposition 1.5 (Co-Existence of Rational and Naïve Agents). Fix a consumer population with a share $\eta > 0$ being naïve local thinkers ($\beta > 1$, $\tilde{\beta} < \beta$) and the remaining share $(1 - \eta) > 0$ being rational ($\beta = 1$). We continue concentrating on interior solutions (regarding the choice of target quality q_{tk} and price p_{tk})

by assuming, throughout, that $p_{\max} \rightarrow \infty$.

Fix any Nash equilibrium. By homogeneity of preferences, naïves and rationals share the same preferences outside stores. We show that, in equilibrium, they also share the same expectations about which product they will purchase at any given store. This implies that both consumer groups will enter the same firm (with probability one if there is one firm that offers the highest surplus in expectation and with strictly positive probability if there are multiple firms that offer the highest surplus in expectation). Consider any firm k . There are two cases: (1) If the firm does not fool, all consumers have correct expectation and thus, expect to purchase the same product. (2) If the firm fools, context-sensitive consumers purchase target t^k , but, by the definition of fooling, there exists some naïve type who expects to purchase some other product $a^k \neq t^k$. Conditional on fooling, profit-maximization implies that $u_{t^k}^k(\beta) = u_{a^k}^k(\beta)$ ((ICC) binds). Because in this case, $u_{t^k}^k(\tilde{\beta}) < u_{a^k}^k(\tilde{\beta})$ for any $\tilde{\beta} < \beta$, all naïve consumers expect to purchase product a^k . Moreover, because $u_{t^k}^k(1) - u_{a^k}^k(1) \Leftrightarrow u_{t^k} < u_{a^k}$, rational consumers are also attracted by the attraction product a^k (which, in comparison to the naïves, they also purchase). It follows that at any point of mutual best response, consumers have identical expectations: There is a unique maximum surplus $\bar{u} > 0$ that both rational and naïve consumers expect to receive and are attracted by. In any equilibrium then, all consumers purchase at the same firms. Moreover, if a firm attracts all consumers of one group, it also attracts all consumers of the other group.

We now consider the best response of some firm k to a given competitor offer conditional on attracting a positive share of consumers. Denote the expected utility that all consumers expect to receive outside of firm k , $\bar{u} > 0$. For ease of notation, we drop the superscript k on all variables of firm k . The firm can either choose *not* to fool, selling some product j at price $p_j = q_j - \bar{u}$ (generating surplus $u_j = \bar{u}$) to all consumers and yielding profit $\pi = p_j - c(q_j)$, or it can choose to fool, in which case the firm sells two different products to naïves (target t) and rationals (attraction product a), yielding profit $\pi = \eta(p_t - c(q_t)) + (1 - \eta)(p_a - c(q_a))$. If the firm fools, profit maximization implies that $u_t^k(\beta) = u_a^k(\beta)$ ((ICC) binds) and $u_a = \bar{u}$ (the participation constraint binds).¹⁷

It is clear that fooling yields higher profit than not fooling. Without fooling, the firm maximizes profit by selling $q_j = q^*$ at $p_j = q^* - \bar{u}$. If the firm fools, it could still attract with a product of the same characteristics, sell it at unchanged

¹⁷Otherwise, the firm could increase the price of the target (1) or the price of the attraction product (2) without affecting demand, violating the profit-maximum.

profit to rationals, while increasing profits on the naïves by inducing them to buy another product at the store (Lemma 1.2). We will now determine the *optimal* fooling strategy, that is, the optimal choice of the attraction product and the target. Assume, w.l.o.g., that the firm offers only two products, the attraction product a and the target t .

We begin with store-wide distortions, that is, for all $\{i, j\} \subseteq J^k$, $\theta_i^k = \theta_j^k = \theta^k \in \{Q, P\}$, and, as a first step, define the optimal choice of (q_a, p_a) and (q_t, p_t) for a given context θ^k .

Assume $\theta^k = Q$. From the two optimality conditions, $u_t^k(\beta) = u_a^k(\beta)$ and $u_a = \bar{u}$, we find $p_t = \beta(q_t - q_a) + p_a$ and $p_a = q_a - \bar{u}$. Profit is

$$\pi(q_t, q_a) = \eta [\beta q_t - (\beta - 1) q_a - c(q_t)] + (1 - \eta) [q_a - c(q_a)] - \bar{u}.$$

First-order conditions $\frac{\partial \pi}{\partial q_t} = 0$ and $\frac{\partial \pi}{\partial q_a} = 0$ yield $c'(q_t) = \beta \Leftrightarrow q_t = q^Q$ and $c'(q_a) = 1 - \frac{\eta}{1-\eta}(\beta - 1)$, respectively. Second-order conditions hold by strict convexity of $c(q)$. Quality q_a so defined is valid if and only if it yields $q_a \geq q_{\min}$, so we have $q_a = \underline{q}_a := \max\{q_{\min}, q|_{c'(q_a)=1-\frac{\eta}{1-\eta}(\beta-1)}\}$. Note that for any positive share of naïves, $\eta > 0$, $q_a < q^* < q_t$ (the firm up-sells naïve consumers). As $\eta \rightarrow 0$, q_a approaches the rational benchmark, $q_a \rightarrow q^*$, from below. Fixing $\theta^k = Q$, we can find equilibrium market prices by setting $\pi = 0$. This yields

$$\begin{aligned} p_a &= \eta c(q^Q) + (1 - \eta) c(q_a) - \eta \cdot \beta (q^Q - q_a) \\ p_t &= \eta c(q^Q) + (1 - \eta) c(q_a) + (1 - \eta) \cdot \beta (q^Q - q_a). \end{aligned}$$

In such an equilibrium, $p_t > c(q_t)$ and $p_a < c(q_a)$ if and only if $\beta q^Q - c(q^Q) > \beta q_a - c(q_a)$, which holds by strict convexity of $c(q)$ and by $q^Q = \arg \max[\beta q - c(q)]$. As $\eta \rightarrow 0$, product-supply for the rational consumers approaches the rational benchmark ($q_a \rightarrow q^*$, $p_a \rightarrow c(q^*)$), while the exploitation of naïve consumers persists ($q_t = q^Q \neq q^*$ and $p_t \rightarrow c(q^*) + \beta(q^Q - q^*) > c(q_t)$).

Assume $\theta^k = P$. Analogously to the case of $\theta^k = Q$, we find

$$\pi(q_t, q_a) = \eta \left[\frac{1}{\beta} \cdot q_t + \left(1 - \frac{1}{\beta}\right) \cdot q_a - c(q_t) \right] + (1 - \eta) [q_a - c(q_a)] - \bar{u}.$$

First-order conditions $\frac{\partial \pi}{\partial q_t} = 0$ and $\frac{\partial \pi}{\partial q_a} = 0$ yield $c'(q_t) = \frac{1}{\beta} \Leftrightarrow q_t = q^P$ and $c'(q_a) = 1 + \frac{\eta}{1-\eta} \cdot \left(1 - \frac{1}{\beta}\right) \Leftrightarrow q_a = \bar{q}_a := q|_{c'(q)=1+\frac{\eta}{1-\eta} \cdot (1-\frac{1}{\beta})}$, respectively. Second-order conditions hold by strict convexity of $c(q)$. Note that for any positive share

of naïves, $\eta > 0$, $q_a > q^* > q_t$ (the firm down-sells naïve consumers). As $\eta \rightarrow 0$, q_a approaches the rational benchmark, $q_a \rightarrow q^*$, from above. Fixing $\theta^k = P$, we can find equilibrium market prices by setting $\pi = 0$. This yields

$$\begin{aligned} p_a &= \eta c(q^P) + (1 - \eta)c(q_a) + \eta \cdot \frac{1}{\beta}(q_a - q^P) \\ p_t &= \eta c(q^P) + (1 - \eta)c(q_a) - (1 - \eta) \cdot \frac{1}{\beta}(q_a - q^P). \end{aligned}$$

In such an equilibrium, $p_t > c(q_t)$ and $p_a < c(q_a)$ if and only if $\frac{1}{\beta} \cdot q^P - c(q^P) > \frac{1}{\beta} \cdot q_a - c(q_a)$, which holds by strict convexity of $c(q)$ and by $q^P = \arg \max [q - \beta c(q)]$. As $\eta \rightarrow 0$, product-supply for the rational consumers approaches the rational benchmark ($q_a \rightarrow q^*$, $p_a \rightarrow c(q^*)$), while the exploitation of naïve consumers persists ($q_t = q^P \neq q^*$ and $p_t \rightarrow c(q^*) - \frac{1}{\beta}(q^* - q^P) > c(q_t)$).

To derive the choice of $\theta^k \in \{Q, P\}$ in equilibrium, fix an equilibrium with $\theta^k = Q$ to find

$$\begin{aligned} \bar{u} = q_{a-k} - p_{a-k} &= \eta \left[[q^Q - c(q^Q)] + (\beta - 1)(q^Q - \underline{q}_a) \right] + (1 - \eta) [q_a - c(\underline{q}_a)] \\ &=: \hat{v}^{(Q,Q)}. \end{aligned}$$

Substitute \bar{u} in the (best response) profit function when choosing the opposite context $\theta^k = P$,

$$\pi^k = \eta \left[\frac{1}{\beta} \cdot q^P + \left(1 - \frac{1}{\beta}\right) \cdot \bar{q}_a - c(q^P) \right] + (1 - \eta) [\bar{q}_a - c(\bar{q}_a)] - \bar{u} =: \hat{v}^{(P,P)} - \hat{v}^{(Q,Q)}.$$

If $\pi^k > 0 \Leftrightarrow \hat{v}^{(Q,Q)} < \hat{v}^{(P,P)}$, equilibrium choice of in-store context is $\theta^k = P$, if $\pi^k < 0 \Leftrightarrow \hat{v}^{(Q,Q)} > \hat{v}^{(P,P)}$, it is $\theta^k = Q$, and in the knife-edge case of $\pi^k = 0 \Leftrightarrow \hat{v}^{(Q,Q)} = \hat{v}^{(P,P)}$, firms may choose either of the two in equilibrium.

Now consider product-specific distortions, that is, the possibility of constructing different distortions for products a and t . Then $(\theta_a^k, \theta_t^k) = (P, Q)$ in the unique best response. Analogously to the case of $\theta^k = Q$, we find

$$\pi(q_t, q_a) = \eta [\beta q_t - (\beta - 1)q_a - \beta \bar{u} - c(q_t)] + (1 - \eta) [q_a - \bar{u} - c(q_a)].$$

First-order conditions $\frac{\partial \pi}{\partial q_t} = 0$ and $\frac{\partial \pi}{\partial q_a} = 0$ yield $c'(q_t) = \beta \Leftrightarrow q_t = q^Q$ and $c'(q_a) = 1 + \frac{\eta}{1-\eta}(\beta - 1) \Leftrightarrow q_a = q|_{c'(q)=1+\frac{\eta}{1-\eta}(\beta-1)}$, respectively. Second-order conditions hold by strict convexity of $c(q)$. Note that for any positive share of naïves,

$\eta > 0$, $q_a > q^*$ and as $\eta \rightarrow 0$, q_a approaches the rational benchmark, $q_a \rightarrow q^*$, from above. Whether the firm up- or down-sells, however, now depends on the share of naïves in the population: If the majority of consumers is rational $\eta \leq \frac{1}{2}$, the firm up-sells ($q_a \leq q_t = q^Q$), and if $\eta > \frac{1}{2}$, it down-sells ($q_a > q_t = q^Q$). We can find equilibrium market prices by setting $\pi = 0$. This yields

$$p_a = [\eta c(q^Q) + (1 - \eta)c(q_a) + \eta \cdot (q_a - \beta q^Q)] \cdot \frac{1}{1 + \eta(\beta - 1)}$$

$$p_t = [\beta \cdot \eta c(q^Q) + \beta \cdot (1 - \eta)c(q_a) - (1 - \eta) \cdot (q_a - \beta q^Q)] \cdot \frac{1}{1 + \eta(\beta - 1)}$$

In such an equilibrium, $p_t > c(q_t)$ and $p_a < c(q_a)$ if and only if $\beta q^Q - c(q^Q) > q_a - \beta c(q_a)$, which holds by strict convexity of $c(q)$ and by $q^Q = \arg \max[\beta q - c(q)]$. As $\eta \rightarrow 0$, product-supply for the rational consumers approaches the rational benchmark ($q_a \rightarrow q^*$, $p_a \rightarrow c(q^*)$), while the exploitation of naïve consumers persists ($q_t = q^Q \neq q^*$ and $p_t \rightarrow \beta c(q^*) + \beta q^Q - q^* > c(q_t)$). \square

Chapter 2

Reputational Discrimination

Author: Arno Appfelstaedt

Abstract: Discrimination can arise in tolerant societies via the “spontaneous” coordination of groups on inefficient social norms that deliver reputational rewards to individuals who restrict their interactions to partners of a certain color. For such norms to be sustainable, information about the color of partners needs to be revealed to other members of the group. This essay shows that competition for interactions can generate incentives for information disclosure. In the presence of competition, discriminatory social norms yield benefits for one group at the expense of the other. Although individuals dislike to discriminate, competition induces them to disclose information about the color of partners in order to gain access to the group that benefits or in order to exclude others from it. Competition also generates incentives for groups to coordinate on a discriminatory norm in the first place.

Keywords: Spontaneous Discrimination, Reputation, Repeated Games, Social Norms, Endogenous Information Disclosure

JEL Codes: D74, D83, C73

2.1 Introduction

Discrimination along observable characteristics such as race or ethnicity—we use “color” as an umbrella term in the following—has been a topic in economics at least since Gary Becker’s seminal work on the *Economics of Discrimination* (Becker, 1957). Economists have studied discrimination primarily in labor markets (Lang and Lehmann, 2012), but also in product markets (Holzer and Ihlanfeldt, 1998), marriage markets (Eeckhout, 2006), and housing markets (Massey and Denton, 1993). The

two prevalent explanations are *taste-based* discrimination (Becker, 1957) and *statistical* discrimination (Arrow, 1973; Phelps, 1972). In both cases, the avoidance of productive interactions with individuals of another color is explained with immediate (in the case of statistical discrimination, expected) payoff effects for the decision maker. This includes theories which assume that immediate payoff effects do not exist inherently, but arise endogenously through game play (for example, Arrow, 1973; Coate and Loury, 1993; Mailath, Samuelson and Shaked, 2000; Ramachandran and Rauh, 2016).

In this essay, we explore a different possibility, which is that discrimination arises from reputational (that is, *intertemporal*) concerns regarding the interaction with a different color. This idea has been formalized by Peski and Szentes (2013) who coin it *spontaneous discrimination*. The game underlying their theory is an infinitely repeated random matching game. Discrimination can arise if groups coordinate on not accepting matches from the other color and on punishing in-group members who violate this rule. The intuition is simple: Consider a population of perfectly tolerant individuals who are identical except for physical color, $c \in \{red, green\}$. Assume that the red group has coordinated on a strategy that rejects red agents who accept green matches. If red individuals value future interactions with their in-group sufficiently, then it is rational for them to not accept green matches, that is, to discriminate. Sequential rationality can be satisfied by assuming that the group does not only ostracize non-discriminators, but also individuals who fail to punish the non-discriminators. Following nomenclature introduced by Kandori (1992), equilibrium coordination that relies on such community enforcement mechanisms may be termed a “social norm”. Spontaneous discrimination is a *discriminatory* social norm.

While the theory of spontaneous discrimination is intellectually attractive, its realism must be judged in the light of the strong requirements it poses on information disclosure: To enable discrimination, information about the color of chosen partners has to become publicly available. Only then can community enforcement work. Importantly, to satisfy sequential rationality and make second-order punishments work, not only does information about the color of immediate partners has to become public knowledge, but also information about the color of the partners of partners, the color of partners of partners of partners, and so on. In the words of Kandori (1992, p.64), “the crux of the matter is information transmission among the community members.” Peski and Szentes (2013) assume that information about the color of partners travels automatically. It seems obvious, however, that such strong

requirements on the extent of public knowledge about (the history of) individual interactions are unlikely to be met if this information is not actively promoted by individuals in society. This essay analyzes whether and under what circumstances such incentives may exist: Can spontaneous discrimination emerge in tolerant societies if information about the color of partners must be voluntarily revealed by individuals themselves or by observing others?

The essay makes three contributions: (1) It offers a simple (belief-free) and tractable theoretical framework to study incentives for information disclosure in the presence of discriminatory social norms. (2) It shows that incentives to reveal information do not exist in a benchmark model that follows assumptions on interactions made by Peski and Szentes (2013). While spontaneous discrimination thus breaks down in the benchmark, (3) the essay continues to show that it can be realistically re-lived by introducing competition for interactions. In this case, not only does the classical result of *out-group* discrimination emerge, but so do social norms of *in-group* discrimination. In the presence of competition, discriminatory social norms yield benefits for one group at the expense of the other. Individuals disclose information about the color of partners in order to gain access to the (preferably small) group that benefits as well as to exclude others from it.

Our model exploits the concept of “social color” introduced by Peski and Szentes (2013). Social color is a dynamic label which allows information about the color of partners to be temporarily attached to the individual. Before deciding whether to interact with someone, individuals observe the respective agent’s physical color $c \in \{red, green\}$ and his social color $s \in \{red, green\}$. This allows us to characterize discriminatory social norms via simple, belief-free (Markov) equilibrium strategies that condition on this information and nothing else. Being conditional on social color, these strategies can deliver intertemporal incentives (punishments or rewards) for discrimination. In the benchmark model, social color reveals information about the color of chosen partners automatically (that is, *exogenously*). We endogenize information disclosure by endogenizing the evolution of social color. In the extended model, social color reveals information about the color of a chosen partner if and only if the decision maker herself *or* another individual who happens to observe the interaction voluntarily decides to reveal this information to the public.

A broad literature on spontaneous discrimination has yet to develop. An early mention of the possibility for such equilibria to emerge features in Dal Bó (2007). Eguia (2015) shows that discriminatory norms can incentivize members of the group

with a relatively disadvantaged status to assimilate, embracing the norms of the more advantaged group. This relates to our finding that under competition, groups may in-group discriminate in order to receive the benefits of the out-group. A different approach to “reputational discrimination” is taken by Choy (2017). Here, individuals build trust among their peers by limiting interactions to the in-group. To our knowledge, ours is the first essay to formally study voluntary incentives for information disclosure in the presence of discriminatory norms.

The focus on information disclosure relates this essay to more general treatments of reporting incentives in repeated games. This literature has historically concentrated on the effect of social norms on welfare-enhancing cooperation in homogeneous groups, typically modeled in the context of a repeated prisoners’ dilemma. An important reference is the earlier cited paper by Kandori (1992). The concept of “social color” can be regarded a specific version of the labeling/stigmatization mechanism studied by this author. Kandori (1992) investigates the minimum degree of information that must be transmitted for social norms to be enforceable, but does not explore incentives for such information transmission. This is done, on the other hand, by Ali and Miller (2016). Because their game does not feature different groups in society, however, they cannot identify reporting incentives that relate to group status and discrimination as our model does.

While in the classical theories of taste-based and statistical discrimination, market competition generally leads to less discrimination (see, for example Darity and Williams, 1985), we are obviously not the first to argue that the opposite prediction might be true. There is a long-standing argument in sociology which asserts this claim, see, for instance, Bobo and Hutchings (1996). When studying inter-group conflict and discrimination, this literature generally regards the ethnic group as a coherent unit, and argues that competition can generate incentives for the group to exclude others from the consumption of a scarce resource. A similar argument is made—using a formal microeconomic model—by Bramoullé and Goyal (2016). We give a short treatise of this issue in the context of our model at the end of section 2.4.

Finally, our model can be seen as one way to rationalize individual incentives to seek group status through out-group discrimination and stigmatization as studied in length by McAdams (1995). In a similar vein it relates to “behavioral” models of group identity and social norms such as Akerlof and Kranton (2000) and Bénabou and Tirole (2012).

The chapter proceeds as follows. The next section presents a benchmark model

with exogenous information disclosure and—in the context of the model—restates the result of “spontaneous discrimination” proposed by Peski and Szentés (2013). Section 2.3 and Section 2.4 extend the benchmark model by introducing endogenous information disclosure and competition for interactions, respectively. Section 2.5 concludes.

2.2 Benchmark: Spontaneous Discrimination

The benchmark captures the essence of Peski and Szentés (2013) in a framework with finitely many agents and deterministic payoffs. Subsequent sections extend the benchmark model to capture endogenous information disclosure (Section 2.3) and competition for interactions (Section 2.4).

2.2.1 The Benchmark Model

Consider a population with individuals of two distinct physical colors, $c \in \{red, green\}$. Each individual $i \in I$, $|I| = n$ finite, belongs to one of the two groups, $c_i \in \{red, green\}$. The number of individuals of color c is denoted n_c , $n_c \geq 3$. For any color $c \in \{red, green\}$, the opposite color is denoted $-c \in \{green, red\}$. Time is discrete and infinite, $t = 0, 1, 2, 3, \dots$. The common discount factor is $\delta < 1$.

Interactions. Payoff is produced in pairwise interactions that involve one principal and one agent.¹ Opportunities for interactions arrive randomly. Each period t , nature uniform randomly draws two individuals from the population—one of which is called the principal $p(t)$ and the other is called the match $\mu(t)$. The probability for any individual i to be the principal is $\frac{1}{n}$. The conditional probability for any individual $j \neq p(t)$ to be the match is $\frac{1}{n-1}$.

The principal observes the match and then decides whether she wants to select him as agent. If she selects him as agent, $a(t) = \mu(t)$, the opportunity realizes. In that case, the interaction produces immediate payoff $H > 0$ of which the principal receives share $\pi > 0$ and the agent receives share $1 - \pi > 0$. Alternatively, the principal can destroy the opportunity, $a(t) = \emptyset$. In that case, no payoff is produced in period t . After the principal has taken a decision, payoffs are realized and the game moves to the next period. Individuals are forward looking and maximize the discounted stream of payoffs generated from current and future interactions.

¹In the framework of Peski and Szentés (2013) the two roles were called employer and worker, respectively.

Social Color. Additional to her *physical* color $c_i \in \{red, green\}$, each individual has a perfectly observable *social* color $s_{i,t} \in \{red, green\}$. We call tuple $\theta_{i,t} := (c_i, s_{i,t})$ the *type* of individual i . The type space is $\Theta := \{red, green\} \times \{red, green\}$.

Social color is a dynamic marker which allows information about the type of agent being chosen to be temporarily attached to the principal. By default, the social color of an individual corresponds to her physical color. In period $t = 0$, $s_{i,t} = c_i$ for all $i \in I$. Moreover, for all $j \neq p(t)$, $s_{j,t+1} = c_j$. As for the principal, the evolution of her social color depends on whether or not she decides to interact with her match. If she destroys the opportunity, $a(t) = \emptyset$, then, in line with the other individuals, $s_{p(t),t+1} = c_{p(t)}$. If, instead, she selects the match as her agent, $a(t) = \mu(t)$, information about the type of agent she interacts with automatically becomes part of her next period social color. In particular, if $a(t) = \mu(t)$, then

$$s_{p(t),t+1} = \begin{cases} c_{a(t)} & \text{with probability } .5, \\ s_{a(t),t} & \text{with probability } .5.^2 \end{cases} \quad (2.1)$$

Because information about the type of agent is automatically attached to the principal, this is a model of *exogenous* information disclosure. Crucially, if $-c_{p(t)} \in \{c_{\mu(t)}, s_{\mu(t),t}\}$, then with positive probability, $s_{p(t),t+1} = -c_{p(t)}$. That is, accepting an agent who carries the *opposite* physical color in his type vector will lead, with some probability, to the principal being temporarily associated with the opposite color as well.

Figure 2.1 summarizes the structure of the game.

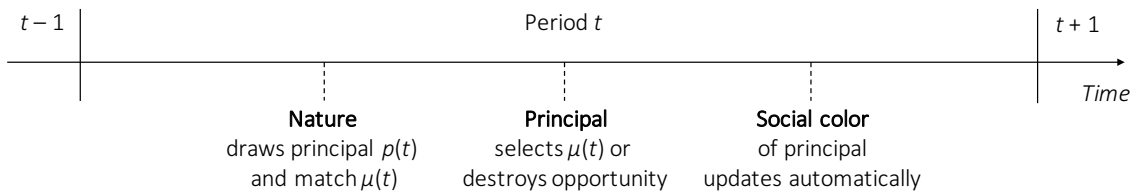


Figure 2.1: Stage Game with Exogenous Information Disclosure (Benchmark)

²We choose probability .5 for technical convenience. The crucial assumption for the equilibria in Proposition 2.1 to emerge is that each option has strictly positive probability.

2.2.2 Equilibrium

Following Peski and Szentes (2013), we are interested in equilibria in which the decision to accept or reject match $\mu(t)$ depends solely on the color of the principal $c_{p(t)} \in \{red, green\}$ as well as on the physical and social color of the match, $(c_{\mu(t)}, s_{\mu(t),t}) \in \{red, green\} \times \{red, green\}$.³ Given our slightly adjusted set-up featuring discrete time and constant payoffs, an adequate equilibrium concept is stationary Markov perfect equilibrium with the Markov state defined by the color of the principal $c_{p(t)}$ and the type of the match $\theta_{\mu(t),t} = (c_{\mu(t)}, s_{\mu(t),t})$.

Social Norms. We characterize the (stationary Markov perfect) equilibrium by characterizing for each group $c \in \{red, green\}$ the set of types $\theta \in \Theta = \{red, green\} \times \{red, green\}$ that the group accepts as agent. We denote this set $A(c)$ and call it the *social norm* of group c . A tuple of social norms $(A(red), A(green))$ maps into equilibrium behavior as follows: If $\theta_{\mu(t),t} \in A(c_{p(t)})$, then $a(t) = \mu(t)$. If $\theta_{\mu(t),t} \notin A(c_{p(t)})$, then $a(t) = \emptyset$. We make the following tie-break assumption: When indifferent whether to accept or reject a match, individuals accept.

The characterization of equilibria in the benchmark model mirrors the main findings from Peski and Szentes (2013) (see their Proposition 1 in particular): While colorblind behavior is always an equilibrium, under certain parameter restrictions, equilibria arise in which group c does not accept agents of the opposite color. That is, they out-group discriminate. This is the outcome referred to by the authors as *spontaneous discrimination*.

Proposition 2.1 (Benchmark: Discrimination with exogenous information disclosure). *In equilibrium, group $c \in \{red, green\}$ follows one of two social norms: A colorblind social norm, $A(c) = \Theta$, or a social norm of out-group discrimination, $A(c) = \{(c, c)\}$. The social norm followed by group c is independent of the social norm followed by the other group, $A(-c)$. The colorblind social norm always exists. The discriminatory social norm $A(c) = \{(c, c)\}$ exists for group $c \in \{red, green\}$ if*

³Allowing the decision to also depend on the social color of the principal, $s_{p(t),t} \in \{red, green\}$, does not change results. Identical to the analysis of Peski and Szentes (2013), the best response of the principal would then be independent of her social color, implying that equilibrium strategies do not depend $s_{p(t),t}$. We simplify notation by not including this (essentially irrelevant) possibility in the definition of the equilibrium.

and only if

$$\pi < .5 \cdot \delta \cdot \frac{n_c - 1}{n(n - 1)} \cdot (1 - \pi). \quad (\text{DC Benchmark})$$

Before formally proving Proposition 2.1, we comment on the incentive structure of the game and give intuition for the result. In comparison to standard models of racial discrimination, discrimination in this model is driven neither by a taste for color nor by statistical payoff differences between colors. This is reflected in the myopic best response, which—regardless of the colors of principal and match—is to be colorblind and select the match as agent. The colorblind group norm $A(c) = \Theta$, which implies that individuals of any type $\theta \in \Theta = \{\text{red}, \text{green}\} \times \{\text{red}, \text{green}\}$ are accepted as agents by group c , implements this myopic best response, does not need to be enforced and therefore always exists. Discrimination, on the other hand, implies that there exists some type $\theta' \in \Theta$ who is not accepted as agent by group c , that is, $\theta' \notin A(c)$. Because discrimination is non-myopic, there must be some element enforcing such a norm in equilibrium. However, there are no exogenous enforcement mechanisms. The enforcing element is the norm itself.

Consider the discriminatory social norm in Proposition 2.1: $A(c) = \{(c, c)\}$. On the equilibrium path group c does not accept agents of the opposite group. The enforcing element of the norm lies *off* the equilibrium path: If an individual of group c deviates and accepts an agent of the opposite group, then with positive probability, her social color will become $s_{i,t+1} = -c$, see equation (2.1). As a result, her type will temporarily become $(c, -c)$. Because $(c, -c) \notin A(c) = \{(c, c)\}$, the social norm of her group contains a punishment for the deviating individual: For as long as she carries the opposite color in her type vector she is ostracized from her group and will be treated as if she was an out-group member. If this punishment is sufficiently painful—in particular, if (DC Benchmark) is satisfied—the individual will not deviate to begin with and $A(c) = \{(c, c)\}$ can be sustained as a social norm. Note that the same mechanism implies that individuals who *fail to punish* (i.e., reject) an agent of type $(c, -c)$ will receive equal punishment. This allows the norm to satisfy sequential rationality (subgame perfectness).

Two characteristics of the equilibrium—which our model shares with Peski and Szentes (2013)—are particularly notable. First, discrimination can emerge as a social norm although there exists no individual in society who benefits from the rejection of agents. This characteristic will lead to a break-down of the norm once we require

endogenous information disclosure (see section 2.3). Second, only discrimination against the *out-group* can be enforced in equilibrium. These findings are not immutable. We show in section 2.4 how competition for interactions can generate both, incentives for information disclosure and social norms of *in-group* discrimination.

In the remainder of this section we talk about the technical details of how to solve for the equilibrium and, subsequently, formally prove Proposition 2.1.

2.2.3 Technical Derivations

We comment on our solution technique. These comments remain valid for the analysis of the extended model in subsequent sections of this essay. Note that stage game payoffs do not depend on private information exogenous to the decision maker (the principal). Because equilibrium strategies defined by social norms ($A(\textit{red}), A(\textit{green})$) also do not condition on private information—recall that types $\theta_{i,t} = (c_i, s_{i,t})$ are common knowledge—, the payoff structure when considering best responses (to social norms) will be belief-free. We therefore do not have to define a belief system. Because the game features a discounted, time-separable payoff structure, we can solve for stationary Markov perfect equilibria using the one-shot deviation principle.⁴

To check whether a tuple of social norms ($A(\textit{red}), A(\textit{green})$) defines an equilibrium, fix the tuple and consider some period t . Exploiting the one-shot deviation principle, we allow decision makers to take *any* action in period t , but constrain their actions in all periods $\tau > t$ to the equilibrium. Actions in period t need therefore to be checked for their profitability with regard to *immediate* payoffs and with regard to consequences they may have for expected *future* payoffs—under the assumption that, in the future, everybody (including the current decision maker) will strictly follow the social norm of his or her group, $A(c_i) \in (A(\textit{red}), A(\textit{green}))$). Immediate payoffs (in period t) depend solely on the employment decision of principal $p(t)$. Expected payoffs regarding any future period $\tau > t$, on the other hand, depend solely on the distribution of types $\theta_{i,\tau}$ in period τ . This is due to the time-invariant random nature of the matching game and equilibrium employment decisions (= future employment decisions) being dependent only on the type of potential principals and agents. The question about how actions in period t can influence future payoffs is thus equivalent to the question about how actions in period t can influence future type-distributions.

⁴We omit a proof of this standard result. Because all payoff-relevant information when deriving the best-response is belief-free, the proof is essentially equivalent to a proof of the one-shot deviation principle when considering subgame perfectness in a repeated game with perfect monitoring, see, for example, Proposition 2.2.1 in Mailath and Samuelson (2006).

Consider the evolution of types from period t to period $t + 1$. By assumption, $s_{j,t+1} = c_j$ for all individuals $j \neq p(t)$, implying that their type in period $t + 1$ is already fixed. The only variable type in period $t + 1$ is that of the current principal, $\theta_{p(t),t+1} = (c_{p(t)}, s_{p(t),t+1})$, whose next period social color $s_{p(t),t+1} \in \{c_{p(t)}, -c_{p(t)}\}$ may depend on period t actions. Period t actions only indirectly impact types beyond $t + 1$: Because future employment decisions follow type-dependent equilibrium strategies, the evolution of types from period τ to period $\tau + 1$, $\tau > t + 1$, depends only on the type-distribution in period τ . The impact on continuation payoffs of period t actions is thus restricted to their impact on the type distribution in period $t + 1$, which is variable only in $\theta_{p(t),t+1} = (c_{p(t)}, s_{p(t),t+1})$: Let $V_i(t)$ denote the continuation payoff of individual i in period t . In any period t , $V_i(t)$ depends solely on the physical color and the next-period social color of the current principal $p(t)$. We write $V_i^{(c,s)}$ for the continuation payoff of individual i if $\theta_{p(t),t+1} = (c, s) \in \Theta$.

Assume that there are two actions a and a' that yield immediate payoff $u_i(a)$ and $u_i(a')$, respectively. Assume further that action a is associated with the principal having social color $s_{p(t),t+1} = s$ and action a' is associated with the principal having social color $s_{p(t),t+1} = s'$. Exploiting the one-shot deviation principle, saying that i prefers action a over a' is then equivalent to the statement $u_i(a) + \delta V_i^{(c,s)} \geq u_i(a') + \delta V_i^{(c,s')}$. We are now ready to prove Proposition 2.1.

Proof of Proposition 2.1. Consider a period t with a principal $p(t)$ of physical color c . Her continuation payoff is $V_{p(t)}^{(c,c)}$ if $s_{p(t),t+1} = c$ and $V_{p(t)}^{(c,-c)}$ if $s_{p(t),t+1} = -c$. Note first that the individual always accepts a match of type $\theta_{\mu(t),t} = (c, c)$: Accepting and rejecting both yield $s_{p(t),t+1} = c$ with probability 1. Accepting, however, yields immediate payoff $\pi H > 0$, while rejecting yields zero immediate payoff. If there exists a type $\theta_{\mu(t),t}$ that the individual rejects it follows that $-c \in \theta_{\mu(t),t}$.

Assume that there exists a type θ' satisfying $-c \in \theta'$ that the principal rejects. It follows that among the types she rejects is type $\theta_{\mu(t),t} = (-c, -c)$: The principal foregoes the immediate payoff from accepting θ' only if rejecting yields higher continuation payoffs, implying $V_{p(t)}^{(c,c)} > V_{p(t)}^{(c,-c)}$. Given $V_{p(t)}^{(c,c)} > V_{p(t)}^{(c,-c)}$, however, accepting type $(-c, -c)$ yields strictly lower expected continuation payoff than accepting types $(c, -c)$ or $(-c, c)$. In particular, accepting type $(-c, -c)$ yields continuation payoff $V_{p(t)}^{(c,-c)}$, while accepting type $(c, -c)$ or $(-c, c)$ yields continuation payoff $\frac{1}{2}V_{p(t)}^{(c,c)} + \frac{1}{2}V_{p(t)}^{(c,-c)}$. Thus, $A(c) \neq \Theta \Rightarrow (-c, -c) \notin A(c)$. Moreover, because types $\theta_{\mu(t),t} = (c, -c)$ and $\theta_{\mu(t),t} = (-c, c)$ are associated with the same continuation payoff for the principal, it follows that $(c, -c) \in A(c) \Leftrightarrow (-c, c) \in A(c)$.

This leaves three possible norms for color c : $A(c) = \Theta$, $A(c) = \{(c, c)\}$, and $A(c) = \Theta \setminus \{(-c, -c)\}$. It is easy to show that the latter of these, $A(c) = \Theta \setminus \{(-c, -c)\}$, does not exist: Fix one of the three norms for the opposite color $-c$ and assume toward a contradiction that $A(c) = \Theta \setminus \{(-c, -c)\}$. For the norm to exist, principal $p(t)$ would have to reject a match of type $(-c, -c)$. Assume first that $A(-c) \in \{\Theta, \{(-c, -c)\}\}$. But then $\{(c, c), (c, -c)\} \subset A(c)$ and $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$, implying that $V_{p(t)}^{(c,c)} = V_{p(t)}^{(c,-c)}$: Because accepting and rejecting a match of type $\theta_{\mu(t),t} = (-c, -c)$ yield identical continuation payoff, the principal will not reject. A contradiction. Assume instead that $A(-c) = \Theta \setminus \{(c, c)\}$. Then $\{(c, c), (c, -c)\} \subset A(c)$, $(c, -c) \in A(-c)$, but $(c, c) \notin A(-c)$. It follows that $V_{p(t)}^{(c,c)} < V_{p(t)}^{(c,-c)}$. Because accepting a match of type $\theta_{\mu(t),t} = (-c, -c)$ yields a higher continuation payoff than rejecting, the principal will not reject. A contradiction. It follows that $A(c) = \Theta \setminus \{(-c, -c)\}$ does not exist.

Two possible norms remain for each group $c \in \{red, green\}$: The colorblind norm $A(c) = \Theta$ and the discriminatory norm $A(c) = \{(c, c)\}$. Note that $A(-c) \in \{\Theta, \{(-c, -c)\}\}$ implies that $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$. The continuation payoff for individuals of color c is independent of which social norm the opposite color follows: Social norms $A(c)$ and $A(-c)$ are independent.

It remains to be shown that the colorblind norm $A(c) = \Theta$ always exists, while the discriminatory norm $A(c) = \{(c, c)\}$ exists if (DC Benchmark) is satisfied. To prove the former, fix $A(-c) \in \{\Theta, \{(-c, -c)\}\}$ and $A(c) = \Theta$. It follows from $\{(c, c), (c, -c)\} \subset A(c)$ and $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$ that $V_{p(t)}^{(c,c)} = V_{p(t)}^{(c,-c)}$. Because continuation payoff does not depend on the evolution of her social color, a principal of color c accepts any type of agent, confirming $A(c) = \Theta$.

Now consider the discriminatory norm $A(c) = \{(c, c)\}$, again fixing $A(-c) \in \{\Theta, \{(-c, -c)\}\}$. It follows from $(c, c) \in A(c)$, $(c, -c) \notin A(c)$ and $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$ that $V_{p(t)}^{(c,c)} > V_{p(t)}^{(c,-c)}$. Social norm $A(c) = \{(c, c)\}$ implies that individual $p(t)$ will be rejected as a match in period $t+1$ by any principal of physical color c if $s_{p(t),t+1} = -c$. The probability that $c_{p(t+1)} = c$, $p(t+1) \neq p(t)$, is $\frac{n_c-1}{n}$. The conditional probability that $\mu(t+1) = p(t)$ is $\frac{1}{n-1}$. The foregone payoff in such a case is $(1-\pi)H$. The expected loss in in period $t+1$ thus calculates to $\frac{n_c-1}{n(n-1)} \cdot (1-\pi)H$. In fact, $V_{p(t)}^{(c,c)} - V_{p(t)}^{(c,-c)} = \frac{n_c-1}{n(n-1)} \cdot (1-\pi)H$: Assuming that individuals follow equilibrium strategies after period t , $A(c) = \{(c, c)\}$ implies that for all individuals of physical color c , $c_i = c$, $s_{i,\tau} = c$ after period $t+1$. Expected payoffs for individual $p(t)$ therefore differ in $s_{p(t),t+1}$ only regarding payoffs in period $t+1$.

The critical case for $A(c) = \{(c, c)\}$ to exist is that the expected loss in continuation payoff is sufficiently large to reject a match of type $\theta_{\mu(t),t} \in \{(c, c), (c, -c)\}$. In this case, $s_{p(t),t+1} = -c$ only realizes with probability $\frac{1}{2}$, while in the case of accepting a match of type $\theta_{\mu(t),t} = (-c, -c)$, it realizes with probability 1. The immediate loss of rejecting a match calculates to πH . It follows that $A(c) = \{(c, c)\}$ exists if and only if $\pi H < \delta \cdot \frac{1}{2} \cdot \frac{n_c-1}{n(n-1)} \cdot (1-\pi)H \Leftrightarrow \pi < \delta \cdot \frac{1}{2} \cdot \frac{n_c-1}{n(n-1)} \cdot (1-\pi)$. This concludes the proof. □

2.3 Endogenous Information Disclosure

Information disclosure lies at the heart of spontaneous discrimination: If social color was not informative about interactions with the other group, social norm $A(c) = \{(c, c)\}$ would fail to exist. Importantly, sequential rationality demands that social color does not only reveal *direct* interactions with the other group (in which case $c_{a(t)} = -c$), but also interactions with individuals of the in-group who interacted with the opposite color, interactions with individuals of the in-group who interacted with individuals of the in-group who interacted with the opposite color, and so on ad infinitum (in all of the these cases, $c_{a(t)} = c$, but $s_{a(t),t} = -c$). It seems obvious that such strong requirements on the extent of public knowledge about (the history of) individual interactions are unlikely to be met if this information is not actively promoted by individuals in society. A natural question therefore to ask is whether there exist incentives for such information disclosure. Can spontaneous discrimination survive if information about the color of agents needs to be endogenously revealed to society?

2.3.1 A Model of Endogenous Information Disclosure

We endogenize information disclosure by assuming that whenever a principal interacts with an agent ($a(t) \neq \emptyset$), her social color updates according to a public message $m(t)$ that is sent by an individual who observes the interaction. We consider two modes of endogenous information disclosure: self-reports and observer-reports.

Self-Reports. Whenever principal $p(t)$ interacts with an agent, $a(t) \neq \emptyset$, the principal can decide to send a message $m(t) \in \{c_{a(t)}, s_{a(t),t}\}$ or decide to stay silent. If she sends a message, her next period social color reveals that message, $s_{p(t),t+1} = m(t)$. If the principal stays silent, or if she destroys the opportunity, $a(t) = \emptyset$, then $m(t) = \emptyset$.

In that case, the next period social color of the principal conforms to the default, $s_{p(t),t+1} = c_{p(t)}$. Figure 2.2 illustrates the stage game.

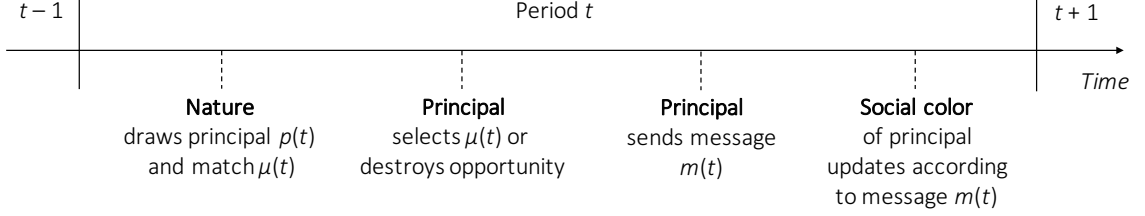


Figure 2.2: Stage Game with Self-Reports

Observer-Reports. Whenever principal $p(t)$ interacts with an agent, $a(t) \neq \emptyset$, nature privately and uniform randomly draws an observer $o(t) \in I^{-p(t)}$. The observer can decide to send a message $m(t) \in \{c_{a(t)}, s_{a(t),t}\}$ or decide to stay silent. If he sends a message, the next period social color of the principal reveals that message, $s_{p(t),t+1} = m(t)$. If the observer stays silent, or if the principal destroys the opportunity, $a(t) = \emptyset$, then $m(t) = \emptyset$. In that case, the next period social color of the principal conforms to the default, $s_{p(t),t+1} = c_{p(t)}$. Figure 2.3 illustrates the stage game.

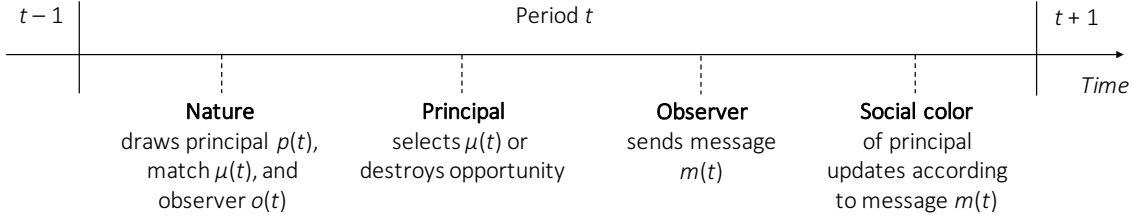


Figure 2.3: Stage Game with Gossip

2.3.2 Equilibrium

We extend the stationary Markov perfect equilibrium concept from the benchmark to include message strategies. Agent choice $a(t) \in \{\mu(t), \emptyset\}$ continues to be determined by social norms $A(c)$, $c \in \{red, green\}$. In the case of self-reports, the Markov state determining message $m(t) \in \{c_{a(t)}, s_{a(t),t}, \emptyset\}$ is given by the type of principal $\theta_{p(t),t} = (c_{p(t)}, s_{p(t),t})$ and the type of agent $\theta_{a(t),t} = (c_{a(t)}, s_{a(t),t})$ she interacts with. In the case of observer-reports, the Markov state determining message $m(t)$ also

includes the type of observer $\theta_{o(t),t} = (c_{o(t)}, s_{o(t),t})$. We make the following tie-break assumption: When indifferent whether to stay silent or send a message, individuals stay silent. Note that a message choice exists in period t if and only if $a(t) \neq \emptyset$, that is, if and only if the principal indeed interacts with an agent. If $a(t) = \emptyset$, then $m(t) = \emptyset$ by default.

We begin with the following observation.

Lemma 2.1 (Messages, if non-empty, inform about interactions with the opposite color). *Fix the physical color of the principal $c_{p(t)} = c \in \{\text{red}, \text{green}\}$ and, in the case of observer-reports, the physical color of the observer $c_{o(t)} \in \{c, -c\}$. In equilibrium,*

1. *If $m(t) \neq \emptyset$, then $m(t) = -c$.*
2. *If $\theta_{a(t),t} = (c, c)$, then $m(t) = \emptyset$.*
3. *If $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, then*
 - (i) *either $m(t) = -c$ for any $\theta_{a(t),t}$,*
 - (ii) *or $m(t) = \emptyset$ for any $\theta_{a(t),t}$.*

Proof. Fix an equilibrium and consider some period t . Individuals send a message, $m(t) \neq \emptyset$, if and only if this yields them strictly higher continuation payoffs than staying silent. On the equilibrium path and after one-shot deviations, continuation payoffs for all individuals are variable only in $\theta_{p(t),t+1}$. Given $c_{p(t)} = c \in \{\text{red}, \text{green}\}$ and, in the case of observer-reports, $c_{o(t)} \in \{c, -c\}$, continuation payoffs for the sender depend solely on $s_{p(t),t+1}$. Sending message $m(t) = c$ yields the same continuation payoff as staying silent (in both cases, $s_{p(t),t+1} = c$). It follows that $m(t) \neq c$, and if $m(t) \neq \emptyset$, then $m(t) = -c$. If $\theta_{a(t),t} = (c, c)$, the only available message is $m(t) = c$. It follows that in this case, $m(t) = \emptyset$. Consider instead the case of $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$. The sender now has the possibility to induce both $s_{p(t),t+1} = -c$ (by sending $m(t) = -c$) and $s_{p(t),t+1} = c$ (by staying silent). If and only if his continuation payoffs are strictly higher if $s_{p(t),t+1} = -c$, he sends $m(t) = -c$ for any type $\theta_{a(t),t} \neq (c, c)$. Otherwise, $m(t) = \emptyset$ for any type $\theta_{a(t),t} \neq (c, c)$. □

Corollary 2.1 (Information about interactions with the opposite color is disclosed with constant probability). *Fix the physical color of the principal $c_{p(t)} = c \in \{\text{red}, \text{green}\}$. In equilibrium, if $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, then*

$m(t) = -c$ with constant probability $\text{Prob}[m = -c] \in [0, 1]$ and $m(t) = \emptyset$ with constant probability $1 - \text{Prob}[m = -c]$.

Consider a principal $p(t)$ of color $c \in \{\text{red}, \text{green}\}$. Intuition for Lemma 2.1 derives from the fact that if $m(t) = \emptyset$, then $s_{p(t), t+1} = c$. It follows that messages can affect future actions and thus, payoffs, only if they associate the principal with the opposite color, that is, only if $m(t) = -c$, implying $s_{p(t), t+1} = -c$. For the sender of the message, the reason why she is able to send message $m(t) = -c$ is irrelevant. In particular, her payoffs do not depend on whether the principal interacted with an agent with physical color $c_{a(t)} = -c$, social color $s_{a(t), t} = -c$, or both. It follows that for any agent who carries $-c$ in his type vector, that is, for any type $\theta_{a(t), t} \in \{(-c, -c), (-c, c), (c, -c)\}$, the message will be identical. Given this insight, Corollary 2.1 is immediate: With self-reports, a principal of color $c_{p(t)} = c$ either sends $m(t) = -c$ for any type $\theta_{a(t), t} \in \{(-c, -c), (-c, c), (c, -c)\}$ (implying constant probability $\text{Prob}[m = -c] = 1$) or for none of these type (implying constant probability $\text{Prob}[m = -c] = 0$). With observer-reports, the same was true if we would fix the physical color of the observer. Because the observer is uniform randomly drawn from $I^{-p(t)}$, the probability for any color $c_{o(t)} \in \{c, -c\}$ is constant. It follows that the probability for $m(t) = -c$ must be constant as well.

The incentive structure of endogenous information disclosure implies that social norms must take one of the forms introduced in Proposition 2.1: A colorblind form, $A(c) = \Theta$, or a self-enforcing form of out-group discrimination, $A(c) = \{(c, c)\}$. With endogenous information disclosure, the existence of the latter hinges on whether there exist individuals who have an incentive to send message $m(t) = -c$ whenever a principal $p(t)$ of color c interacts with an agent of type $\theta_{a(t), t} \in \{(-c, -c), (-c, c), (c, -c)\}$. Intuitively, this criterion must fail: Sending $m(t) = -c$ leads to a punishment of individual $p(t)$ by banning her (temporarily) from interactions with group c . Clearly, the principal does not self-report. In the current setup, the ostracism of individual $p(t)$ also fails to benefit any other individual. In fact, the prospect of productive opportunities being destroyed as a punishment of $p(t)$ must mean that other members of group c are strictly worse off. As a result, observers remain silent as well. We prove this intuition in the Proposition below.

Proposition 2.2 (Discrimination breaks down in the benchmark). *If information disclosure is endogenous, the unique social norm for any group $c \in \{\text{red}, \text{green}\}$ is the colorblind social norm, $A(c) = \Theta$.*

Proof. Consider a principal $p(t)$ of physical color $c \in \{\text{red}, \text{green}\}$. Accepting an agent of type (c, c) yields the same social color and continuation payoff as rejecting him. In equilibrium, a match of type (c, c) will therefore always be accepted, implying $(c, c) \in A(c)$. If there exists $\theta \notin A(c)$, then $\theta \neq (c, c) \Leftrightarrow \theta \in \{(-c, -c), (-c, c), (c, -c)\}$. By Corollary 2.1, accepting an agent of type $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$ yields $m(t) = -c$ with constant probability. It follows that all agents of type $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$ yield the same continuation payoff for the principal. In equilibrium then, either all must be rejected, implying $A(c) = \{(c, c)\}$, or all must be accepted, implying $A(c) = \Theta$. The colorblind social norm $A(c) = \Theta$ always exists. The proof can be made identically to the case with exogenous information disclosure (see the proof of Proposition 2.1).

We show that with endogenous information disclosure, the discriminatory norm $A(c) = \{(c, c)\}$ fails to exist. Fix $A(-c) \in \{\Theta, \{(-c, -c)\}\}$ and $A(c) = \{(c, c)\}$. Then $V_{p(t)}^{(c,c)} > V_{p(t)}^{(c,-c)}$ (see the proof of Proposition 2.1). Clearly, the norm cannot be sustained with self-reports: Principal $p(t)$ can guarantee herself continuation payoff $V_{p(t)}^{(c,c)}$ by staying silent (yielding $s_{p(t),t+1} = c$) after accepting a match of type $\theta_{\mu(t),t} \neq (c, c)$. It follows that $a(t) = \mu(t)$ and $m(t) = \emptyset$ for any type $\theta_{\mu(t),t}$: The norm fails to exist. The norm can also not be sustained with observer-reports. Assume that $a(t) = \mu(t)$ and $\theta_{\mu(t),t} \neq (c, c)$. If $c_{o(t)} = c$, the observer has a strict incentive to stay silent: Exploiting the one-shot deviation principle, $m(t) = -c$ implies that the opportunity for a productive interaction will be destroyed if individual $p(t)$ is matched to a principal of color c in period $t + 1$. This yields a loss $\pi H > 0$ to the principal. With positive probability, this principal is the observer. If $c_{o(t)} = -c$, the observer is indifferent between sending $m(t) = -c$ and staying silent and thus, stays silent: Exploiting the one-shot deviation principle, $A(-c) \in \{\Theta, \{(-c, -c)\}\}$ implies that expected continuation payoffs for the observer are unaffected by $s_{p(t),t+1}$. If $A(-c) = \Theta$, individual $p(t)$ will be accepted as agent irrespective of her social color. If $A(-c) = \{(-c, -c)\}$, individual $p(t)$ will be rejected irrespective of her social color. Because $m(t) = \emptyset$ for any choice $a(t)$, the principal accepts the match irrespective of type $\theta_{\mu(t),t}$: The norm fails to exist. □

2.4 Competition for Interactions

Does the breakdown of spontaneous discrimination imply that the concept is unsuited to explain discrimination in environments that require endogenous information disclosure? This chapter argues against this interpretation. Instead, it offers a theory regarding the nature of interactions that can realistically revive the discriminatory outcome. This theory is based on the idea that discriminatory social norms are more likely to emerge in societies in which agents compete for interactions.

2.4.1 A Model With Competitive Interactions

We introduce competition for interactions by assuming that the principal—additional to being able to select match $\mu(t)$ as her agent—can also select any other individual $j \in I^{-p(t)}$, $j \neq \mu(t)$, to realize the opportunity. In particular, we keep to the existing setup, but assume that in any period t , the choice set of the principal is $a(t) \in I^{-p(t)} \cup \{\emptyset\}$. As before, choice $a(t) = \emptyset$ implies that the principal destroys the opportunity to produce output in period t , while selecting $\mu(t)$ as agent yields output H . Selecting $j \neq \mu(t)$ over $\mu(t)$ is costly: When $a(t) = j \neq \mu(t)$, the pair produces output $L > 0$, $L < H$. Keeping to the benchmark, any output produced is shared between the principal and the agent, with $\pi > 0$ denoting the share going to the principal.

Assumption $L < H$ has two different interpretations: One may think of $\mu(t)$ being an expert for the opportunity that arose in period t —with other agents being less productive on the job. One may also think of all agents having the same productivity, but search for a partner other than the initial match $\mu(t)$ being costly.⁵ Interactions are competitive on the side of potential agents: While the principal always prefers to select $\mu(t)$, any $j \neq \mu(t)$ would prefer her to reject $\mu(t)$ and select him instead. The ratio $\frac{L}{H} \in (0, 1)$ is a measure of the substitutability of $\mu(t)$ and of the degree of competition for jobs. Figure 2.4 illustrates the new sequence of events in period t .

⁵By taking $L \rightarrow H$ we are able to study the marginal case in which the costs of selecting $j \neq \mu(t)$ over $\mu(t)$ go to zero.

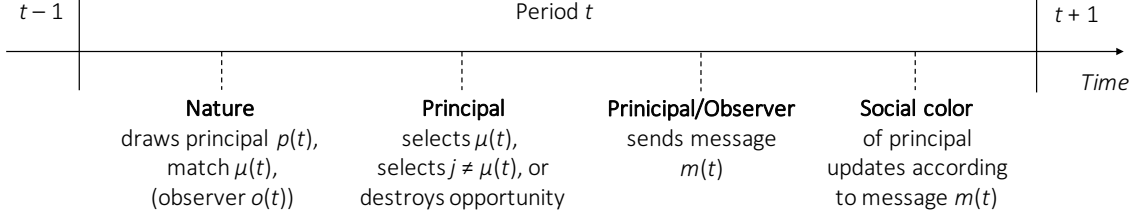


Figure 2.4: Stage Game with Competition (and Endogenous Information Disclosure)

2.4.2 Equilibrium

We continue concentrating on stationary Markov perfect equilibria that allow for a characterization of agent choice via social norms $A(c)$. Given a profile of social norms $(A(\text{red}), A(\text{green}))$ and a draw of principal $p(t)$, let $J^A(t)$ be the period t set of agents who are compatible with the norm:

$$J^A(t) := \{j \in I^{-p(t)} \mid \theta_{j,t} \in A(c_{p(t)})\}.$$

Social norms $A(c)$ then map into the choice of the principal as follows: If $\mu(t) \in J^A(t)$, then $a(t) = \mu(t)$. If $\mu(t) \notin J^A(t)$, then $a(t) = j \in J^A(t)$, each $j \in J^A(t)$ with equal probability. And if $J^A(t) = \emptyset$, then $a(t) = \emptyset$. We keep to the model of endogenous information disclosure (self-reports and observer-reports) introduced in the previous section. As assumed in said section, message choice $m(t)$ in equilibrium follows Markov strategies that condition on the type of principal $\theta_{p(t),t}$, the type of agent $\theta_{a(t),t}$, and, additionally, in the case of observer-reports, the type of observer $\theta_{o(t),t}$.

We begin with the following observations.

Lemma 2.2 (Observations on the incentives for information disclosure). *Lemma 2.1 and Corollary 2.1 (Section 2.3) remain valid.*

Proof. Omitted. (The proof of Lemma 2.1 remains valid.)

□

Proposition 2.3 (Equilibrium candidates). *In equilibrium, group $c \in \{\text{red}, \text{green}\}$ follows one of three social norms: A colorblind social norm, $A(c) = \Theta$, a social norm of out-group discrimination, $A(c) = \{(c, c)\}$, or a social norm of in-group discrimination/out-group favoritism, $A(c) = \Theta \setminus \{(c, c)\}$. The social norm followed by group c is independent of the social norm followed by the other group, $A(-c)$. The colorblind social norm always exists.*

Proof. Consider a principal $p(t)$ of physical color $c \in \{\text{red}, \text{green}\}$. By Corollary 2.1, accepting any agent with type $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$ yields the same expected continuation payoff for the principal. It follows that, in equilibrium, either all types $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$ are accepted as agent, implying $\{(-c, -c), (-c, c), (c, -c)\} \subseteq A(c)$, or all types $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$ are rejected as agent, implying $\{(-c, -c), (-c, c), (c, -c)\} \cap A(c) = \emptyset$. Additionally, social norms may differ in whether type $(c, c) \in A(c)$. This leaves three possible norms: (1) $(c, c) \in A(c)$ and $\{(-c, -c), (-c, c), (c, -c)\} \subseteq A(c) \Leftrightarrow A(c) = \Theta$, (2) $(c, c) \in A(c)$ but $\{(-c, -c), (-c, c), (c, -c)\} \cap A(c) = \emptyset \Leftrightarrow A(c) = \{(c, c)\}$, (3) $(c, c) \notin A(c)$ but $\{(-c, -c), (-c, c), (c, -c)\} \subseteq A(c) \Leftrightarrow A(c) = \Theta \setminus \{(c, c)\}$.

Fix $A(-c)$ to one of these norms. Note that $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$: On the equilibrium path and after one-shot deviations, the continuation payoff for a principal of color c is independent of social norm $A(-c)$. It follows that social norm $A(c)$ is independent of social norm $A(-c)$. Continue to fix $A(-c)$ and assume that group c follows a colorblind norm, $A(c) = \Theta$. Note that in that case, $(c, c) \in A(c) \Leftrightarrow (c, -c) \in A(c)$. Exploiting the one-shot deviation principle, continuation payoffs for principal $p(t)$ are then unaffected by the evolution of her social color and, more generally, by her actions in period t . The payoff-maximizing choice is to comply with norm $A(c) = \Theta$ and accept match $\mu(t)$ irrespective of type $\theta_{\mu(t),t}$. It follows that social norm $A(c) = \Theta$ always exists. □

Lemma 2.3 (Technical observation on continuation payoffs after one-shot deviations). *Fix an equilibrium and consider some period t with $c_{p(t)} = c$. Assuming that deviations in period t are one-shot, then for any individual $i \in I$, continuation payoffs $V_i^{(c,c)}$ and $V_i^{(c,-c)}$ only differ in period $t + 1$.*

Proof. By Proposition 2.3, $A(c) \in \{\Theta, \{(c, c)\}, \Theta \setminus \{(c, c)\}\}$. Fix a tuple of social norms ($A(\text{red}), A(\text{green})$) and consider some period t . Assuming that any deviation in period t is one-shot, expected payoff in any period $\tau > t$ only depends on the type distribution in period τ . Moreover, the type distribution in period $\tau + 1$ depends only on $m(\tau)$. We show that $m(\tau)$ does not depend on $m(\tau - 1)$. It follows that expected payoffs in any period $\hat{\tau} > t + 1$ do not depend on message $m(t) = s_{p(t),t+1}$ and thus, $V_i^{(c,c)}$ and $V_i^{(c,-c)}$ only differ in $t + 1$.

To simplify notation, let $c_{p(\tau)} = c$. We first show that if $A(c) = \Theta$, then $m(\tau) = \emptyset$ and thus, $m(\tau)$ does not depend on $m(\tau - 1)$. Note first that by $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$, expected payoffs for any individual $i \in I$ in

any period $\hat{\tau} > \tau$ conditional on $c_{p(\hat{\tau})} = -c$ do not depend on $m(\hat{\tau} - 1)$ if $c_{p(\hat{\tau}-1)} = c$. If $A(c) \in \Theta$, expected payoffs in any period $\hat{\tau} > \tau$ also do not depend on $m(\hat{\tau} - 1)$ if $c_{p(\hat{\tau})} = c$. It follows that $V_i^{(c,c)} = V_i^{(c,-c)}$ for any individual $i \in I$, and thus $m(\tau) = \emptyset$.

Now consider the case of $A(c) \neq \Theta$. By Lemma 2.1, $m(\tau) = \emptyset$ if $\theta_{a(\tau),\tau} = (c, c)$. Moreover, if $\theta_{a(\tau),\tau} \neq (c, c)$, then $m(\tau) = -c$ with constant probability, and with residual probability, $m(\tau) = \emptyset$. Thus, $m(\tau)$ only depends only on $Prob[\theta_{a(\tau),\tau} \neq (c, c)]$. We show that $Prob[\theta_{a(\tau),\tau} \neq (c, c)]$ does not depend on $m(\tau - 1)$. Assume that $A(c_{p(t)}) = \{(c, c)\}$. Then $Prob[\theta_{a(\tau),\tau} \neq (c, c)] = 0$, irrespective of $m(\tau)$. Assume that $A(c) = \Theta \setminus \{(c, c)\}$. Then $Prob[\theta_{a(\tau),\tau} \neq (c, c)] = 1$, irrespective of $m(\tau)$. It follows that for any $p(\tau) \in I$, $m(\tau)$ does not depend on $m(\tau - 1)$. □

2.4.3 Out-Group Discrimination

When interactions are competitive, compliance with social norm $A(c) = \{(c, c)\}$ implies the following choice of agent: If match $\mu(t)$ is of type (c, c) , the principal selects the match. Otherwise, the principal selects another agent of type (c, c) , each with equal probability.⁶ Because the principal can comply with the norm without having to destroy productive opportunities, not only do her compliance cost decrease, there now exists individuals—those of type $\theta_{j,t} = (c, c)$ —who benefit from the discrimination of others. Importantly, the individual benefit for any agent j of type (c, c) is larger the smaller the set of others who share his type. Intuitively, given that there is competition for interactions, the probability for agent j to interact with the principal is higher the fewer the number of other agents the principal can select. The Lemma below shows that this mechanism can generate incentives for information disclosure: Under a social norm of out-group discrimination, if competition is sufficiently strong, individuals of color c gain from ostracizing others from their group by sending messages about their interactions with the opposite color.

Lemma 2.4 (Information disclosure under a social norm of out-group discrimination). *Fix $A(c) = \{(c, c)\}$ and consider a principal $p(t)$ of color $c_{p(t)} = c$. In equilibrium, if the principal selects an agent of type $\theta_{a(t),t} = (c, c)$, then $m(t) = \emptyset$. If the principal selects an agent of type $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, then*

⁶Under social norm $A(c) = \{(c, c)\}$, the set of norm-compatible agents $J^A(t)$ is always non-empty. Because $s_{k,t} = c_k$ for all $k \neq p(t-1)$, $J^A(t)$ either includes all individuals $j \neq p(t)$ of color c (then $|J^A(t)| = n_c - 1 > 0$) or includes all individuals $j \neq p(t)$ of color c *except* the principal of the previous period (then $|J^A(t)| = n_c - 2 > 0$).

1. *The principal will not self-report: With self-reports, $m(t) = \emptyset$.*
2. *Observers of the opposite color will not report: If $c_{o(t)} = -c$, then $m(t) = \emptyset$.*
3. *Observers of the same color will report if and only if competition ($\frac{L}{H}$) is sufficiently high: If $c_{o(t)} = c$, then $m(t) = -c$ if*

$$\frac{L}{H} > \frac{\pi}{1 + \frac{n-c}{n_c-1}(1-\pi)}, \quad (\text{IC Out-Group})$$

and $m(t) = \emptyset$ otherwise.

Proof. Fix $A(c) = \{(c, c)\}$. Fix $A(-c)$ to one of three norms defined in Proposition 2.3. By Lemma 2.1, if the principal selects an agent of type $\theta_{a(t),t} = (c, c)$, then $m(t) = \emptyset$. Assume for the rest of the proof that the principal selects an agent of type $\theta_{a(t),t} \neq (c, c)$. Immediate payoffs do not depend on message $m(t)$. Assuming that any deviation in period t is one-shot, future payoffs are affected by message $m(t)$ only in period $t + 1$ (see Lemma 2.3). Social norm $A(c) = \{(c, c)\}$ implies that if $m(t) = -c$, individual $p(t)$ will be rejected by group c in period $t + 1$ and replaced by an alternative agent of type (c, c) . If $m(t) = \emptyset$, she will be accepted. Because $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$, the probability that $p(t)$ is accepted by the opposite group $-c$ is, on the other hand, not affected by $m(t)$.

For the principal as well as for observers of the opposite color, $c_{o(t)} = -c$, the incentives to disclose information are unchanged from the non-competitive benchmark (see Section 2.3). For the principal, $m(t) = -c$ implies an expected loss in period $t + 1$. For an observer of the opposite color, $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$ implies that $m(t) = -c$ yields the same continuation payoffs as $m(t) = \emptyset$. It follows that in the case of self-reports as well as in the case of $c_{o(t)} = -c$, there is no information disclosure, that is, $m(t) = \emptyset$. This proves parts 1 and 2 of the Lemma.

Competition affects the incentives of observers who are of the same color as the principal, $c_{o(t)} = c$. Sending $m(t) = -c$ now carries a negative and a positive effect. With positive probability ($\frac{1}{n} \cdot \frac{1}{n-1}$), the observer is principal in period $t + 1$ and matched to individual $p(t)$. In that case he incurs a loss of $\pi(H - L)$ because he will have to reject $p(t)$ and select an alternative agent of type (c, c) . This is the negative effect. The positive effect is due to the fact that other individuals of group c will also have to reject $p(t)$. Because $\theta_{o(t),t+1} = (c, c)$, this increases the chances of the observer to be selected as an alternative agent in period $t + 1$.

Assume that $c_{p(t+1)} = c$, $p(t+1) \notin \{p(t), o(t)\}$ (probability $\frac{n_c-2}{n}$). There are two mechanisms by which individual $o(t)$ will benefit from having sent $m(t) = -c$. The first mechanism is that $m(t) = -c$ increases the number of realization of $\mu(t+1)$ for which $p(t+1)$ needs to select an alternative agent. In particular, if $\mu(t+1) = p(t)$ (conditional probability $\frac{1}{n-1}$), message $m(t) = -c$ induces the principal to reject $p(t)$ and select an alternative agent of type $\theta_{j,t+1} = (c, c)$. With positive probability ($\frac{1}{n_c-2}$) he will select $o(t)$.⁷ The second mechanism is that $m(t) = -c$ reduces the number of alternative agents whom the principal can select whenever else he has to reject $\mu(t+1)$. In particular, if $c_{\mu(t+1)} = -c$ (conditional probability $\frac{n-c}{n-1}$), message $m(t) = -c$ yields the observer a probability of $\frac{1}{n_c-2}$ instead of $\frac{1}{n_c-1}$ to be selected.⁸ Whenever the observer is selected as alternative agent, this yields him a payoff of $(1 - \pi)L$. The total expected gain from sending $m(t) = -c$ thus sums up to $\frac{n_c-2}{n} \left[\frac{1}{n-1} \cdot \frac{1}{n_c-2} + \frac{n-c}{n-1} \left(\frac{1}{n_c-2} - \frac{1}{n_c-1} \right) \right] (1 - \pi)L$.

The observer sends $m(t) = -c$ if the expected gains in period $t+1$ strictly overweigh the expected loss, that is, if

$$\begin{aligned} \frac{1}{n} \cdot \frac{1}{n-1} \cdot \pi(H - L) &< \frac{n_c-2}{n} \left[\frac{1}{n-1} \cdot \frac{1}{n_c-2} + \frac{n-c}{n-1} \left(\frac{1}{n_c-2} - \frac{1}{n_c-1} \right) \right] (1 - \pi)L \\ \Leftrightarrow \frac{L}{H} &> \frac{\pi}{1 + \frac{n-c}{n_c-1}(1 - \pi)}. \end{aligned}$$

Otherwise, the observer will remain silent, $m(t) = \emptyset$. This concludes the proof of part 3 of the Lemma. \square

Competition also affects the incentives of the principal to comply with norm $A(c) = \{c, c\}$: The possibility to select an alternative agent of type (c, c) instead of having to destroy the productive opportunity if $\theta_{\mu(t),t} \neq (c, c)$ lowers her compliance cost from πH to $\pi(H - L)$. Moreover, being punished for violating the norm now comes at a greater cost: If $s_{p(t),t+1} = -c$, the principal does not only lose the opportunity to be selected as a match by group c , she also loses the opportunity to be selected as an alternative agent in the case that principals of her group need to

⁷If $m(t) = -c$, the set of norm-compatible agents $J^A(t+1) = \{j \in I^{-p(t+1)} \mid \theta_{j,t+1} = (c, c)\}$ includes all individuals of color c except the current principal $p(t+1)$ and the previous principal $p(t)$. Because the alternative agent $j \neq \mu(t+1)$ is selected with uniform probability from the set $J^A(t+1)$, the probability for $o(t)$ to be selected is $\frac{1}{|J^A(t+1)|} = \frac{1}{n_c-2}$.

⁸If $m(t) = \emptyset$, the set of norm-compatible agents $J^A(t+1) = \{j \in I^{-p(t+1)} \mid \theta_{j,t+1} = (c, c)\}$ includes individual $p(t)$. The probability for $o(t)$ to be selected is then $\frac{1}{|J^A(t+1)|} = \frac{1}{n_c-1}$.

replace a match of the opposite color. The following Lemma answers the question under which conditions a principal indeed complies with social norm $A(c) = \{c, c\}$.

Lemma 2.5 (Agent choice under a social norm of out-group discrimination). *Fix $A(c) = \{(c, c)\}$ and consider a principal $p(t)$ of color $c_{p(t)} = c$. If the match is of type $\theta_{\mu(t),t} = (c, c)$, then $a(t) = \mu(t)$. If the match is of type $\theta_{\mu(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, then the principal complies with the norm by selecting an alternative agent $j \neq \mu(t)$ of type $\theta_{j,t} = (c, c)$ if*

$$\pi \left(1 - \frac{L}{H}\right) < \text{Prob}[m = -c] \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot \left(1 + \frac{n_c - 1}{n_c - 1} \cdot \frac{L}{H}\right) (1 - \pi),$$

(DC Out-Group)

and selects the match, $a(t) = \mu(t)$, (does not comply) otherwise.

Proof. Fix $A(c) = \{(c, c)\}$. Fix $A(-c)$ to one of three norms defined in Proposition 2.3. If the principal selects match $\mu(t)$, she receives immediate payoff πH . If she selects any other agent, she receives immediate payoff $\pi L < \pi H$. Exploiting the one-shot deviation principle, continuation payoffs depend only on $s_{p(t),t+1} \in \{c, -c\}$ and are affected only in period $t + 1$ (see Lemma 2.3). Social norm $A(c) = \{(c, c)\}$ implies that if $s_{p(t),t+1} = -c$, individual $p(t)$ will be rejected by group c in period $t + 1$, while if $s_{p(t),t+1} = c$, she will be accepted. Because $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$, the probability that $p(t)$ is accepted by the opposite group $-c$ is, on the other hand, not affected by $s_{p(t),t+1}$. It follows that $V_{p(t)}^{(c,c)} > V_{p(t)}^{(c,-c)}$.

The expected loss in continuation payoffs if $s_{p(t),t+1} = -c$ amounts to

$$V_{p(t)}^{(c,c)} - V_{p(t)}^{(c,-c)} = \frac{n_c - 1}{n} \cdot \frac{1}{n-1} \cdot (1 - \pi)H + \frac{n_c - 1}{n} \cdot \frac{n_c}{n-1} \cdot \frac{1}{n_c - 1} \cdot (1 - \pi)L.$$

The first term is identical with $V_{p(t)}^{(c,c)} - V_{p(t)}^{(c,-c)}$ in the non-competitive benchmark: It accounts for lost payoffs in the case where $p(t)$ is rejected as match $\mu(t + 1)$: With probability $\frac{n_c - 1}{n}$, an individual $j \neq p(t)$ of group c is principal in period $t + 1$, and with conditional probability $\frac{1}{n-1}$, individual $p(t)$ is her match. If $s_{p(t),t+1} = -c$, individual $p(t)$ will be rejected, resulting in a loss of $(1 - \pi)H$. When interactions are competitive, there is a second source of income that is affected by $s_{p(t),t+1}$: If another individual of group c is principal and her match is of the opposite color, $c_{\mu(t+1)} = -c$ (conditional probability $\frac{n_c}{n-1}$), the principal will select an alternative agent of type (c, c) . If $s_{p(t),t+1} = c$, individual $p(t)$ will be selected with positive probability $\frac{1}{n_c - 1}$.⁹

⁹If $s_{p(t),t+1} = c$, the set of norm-compatible agents $J^A(t + 1) = \{j \in I^{-p(t+1)} \mid \theta_{j,t+1} = (c, c)\}$

In that case, individual $p(t)$ earns payoff $(1 - \pi)L$. If $s_{p(t),t+1} = -c$, individual $p(t)$ will be selected with probability zero. This effect is covered by the second term.

We are ready to determine the one-shot payoff-maximizing action of principal $p(t)$ conditional on match $\mu(t)$. Consider the case of $\theta_{\mu(t),t} = (c, c)$. By Lemma 2.1, if $a(t) = \mu(t)$, then $m(t) = \emptyset$ and thus, $s_{p(t),t+1} = c$. Choice $a(t) = \mu(t)$ maximizes both immediate and continuation payoffs. It follows that $a(t) = \mu(t)$. Consider the case of $\theta_{\mu(t),t} \neq (c, c)$. If the principal violates the norm and selects the match, she receives immediate payoff πH . The continuation payoff associated with this action is $V_{p(t)}^{(c,-c)}$ with $Prob[m = -c]$ and $V_{p(t)}^{(c,c)}$ with residual probability (see Corollary 2.1). Alternatively, the principal can follow the norm and select some $j \neq \mu(t)$ of type (c, c) , yielding (lower) immediate payoff πL , but continuation payoff $V_{p(t)}^{(c,c)}$ with probability 1. This is strictly payoff-maximizing if

$$\begin{aligned} \pi(H - L) &< Prob[m = -c] \cdot \delta \cdot [V_{p(t)}^{(c,c)} - V_{p(t)}^{(c,-c)}] \\ \Leftrightarrow \pi \left(1 - \frac{L}{H}\right) &< Prob[m = -c] \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \left(1 + \frac{n-c}{n_c - 1} \cdot \frac{L}{H}\right) (1 - \pi). \end{aligned}$$

Otherwise, the payoff-maximizing choice is to select the match, $a(t) = \mu(t)$. □

Equipped with Lemmas 2.4 and 2.5, we are ready to characterize the conditions for out-group discrimination to be enforceable: Social norm $A(c) = \{(c, c)\}$ can be enforced if norm violations are reported with positive probability (Lemma 2.4) and if the resulting punishments are sufficiently strong to deter individuals from violating the norm (Lemma 2.5). Clearly, the norm cannot be sustained with self-reports as principals themselves do not have incentives to disclose their interactions with agents of the opposite color. With observer-reports, enforceability depends critically on the level of competition ($\frac{L}{H}$). If competition is sufficiently high such that the information disclosure constraint (IC Out-Group) is satisfied, observers of color c have incentives to ostracize group members who violate the norm. Given that (IC Out-Group) is satisfied, the probability that a principal of group c is reported when interacting with the opposite color is then equal to the probability that the principal is observed by a member of her own group, that is,

$$Prob[m = -c] = Prob \left[c_{o(t)} = c \mid c_{p(t)} = c \right].$$

includes individual $p(t)$. The probability for $p(t)$ to be selected is then $\frac{1}{|J^A(t+1)|} = \frac{1}{n_c - 1}$.

By the assumption that the observer is uniform randomly drawn from $I^{-p(t)}$ this probability equals $\frac{n_c-1}{n-1}$ in any period t . We can thus conclude:

Proposition 2.4 (Out-group discrimination with endogenous information disclosure). *If interactions are competitive, a social norm of out-group discrimination $A(c) = \{(c, c)\}$ may exist. The norm can be sustained with observer-reports, but not with self-reports. In particular, the norm exists for group $c \in \{\text{red}, \text{green}\}$ if and only if*

$$\pi \left(1 - \frac{L}{H}\right) < \text{Prob}[m = -c] \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot \left(1 + \frac{n_{-c}}{n_c - 1} \cdot \frac{L}{H}\right) (1 - \pi),$$

(DC Out-Group)

where $\text{Prob}[m = -c] = 0$ (the norm does not exist) if information is disclosed through self-reports, and

$$\text{Prob}[m = -c] = \begin{cases} \frac{n_c - 1}{n - 1} & \text{if } \frac{L}{H} > \frac{\pi}{1 + \frac{n_{-c}}{n_c - 1} (1 - \pi)}, \\ 0 \text{ (the norm does not exist)} & \text{otherwise,} \end{cases}$$

if information is disclosed through observers.

Proof. For given probability $\text{Prob}[m = -c]$, the social norm exists if the discrimination constraint (DC Out-Group) is satisfied (Lemma 2.5). With self-reports, $\text{Prob}[m = -c] = 0$ (see Lemma 2.4). Consider observer-reports. Then in any period t , $\text{Prob}[c_{o(t)} = c \mid c_{p(t)} = c] = \frac{n_c-1}{n-1}$. It then follows from Lemma 2.4 that if (IC Out-Group) holds, $\text{Prob}[m = -c] = \frac{n_c-1}{n-1}$, and $\text{Prob}[m = -c] = 0$ otherwise. \square

Compare this result with Proposition 2.1: In the non-competitive benchmark, spontaneous discrimination requires π to be sufficiently small. When interactions are competitive and information can be disclosed by observers, on the other hand, there exists a level of competition $\frac{L}{H}$ such that spontaneous discrimination is a social norm for *any* size of π . Intuitively, competition makes out-group discrimination easier to enforce for two reasons: (1) It makes it easier to substitute out-group members for in-group members and thus, makes discrimination cheaper. (2) Because in-group members benefit from the discrimination of their peers, it generates incentives for observers to enforce the norm.

2.4.4 In-Group Discrimination/Out-Group Favoritism

The competitive environment allows for a discriminatory norm that has so far been nonexistent: Social norm $A(c) = \Theta \setminus \{(c, c)\}$ asks principals of color c to reject any agent who does *not* carry the opposite color $-c$ in their type vector.¹⁰ Because this means that any agent j of color c who carries his default social color $s_{j,t} = c$ will be rejected, this is a form of in-group discrimination. More precisely, social norm $A(c) = \Theta \setminus \{(c, c)\}$ is associated with the following equilibrium agent choice: If the match is of type $\theta_{\mu(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, then the principal selects the match. If the match is of type $\theta_{\mu(t),t} = (c, c)$, then the principal selects an alternative agent j of type $\theta_{j,t} \in \{(-c, -c), (-c, c), (c, -c)\}$, each with equal probability.

We begin by studying incentives for information disclosure. Note that the norm strongly favors members of the opposite group. Agents who carry $-c$ as their physical color, $c_j = -c$, will always be accepted by group c . Agents of physical color c , on the other hand, have to be “socially” associated with the opposite color, $s_{i,t} = -c$, in order to be accepted by their peers. Under such a norm, individuals of group c clearly have a strong incentive to send message $m(t) = -c$. Competition implies that individuals of the opposite group have the opposite incentive: In order to keep the benefits of out-group favoritism to themselves, they have a strict incentive to not send any message that can help individuals of group c to associate with their color.

Lemma 2.6 (Information disclosure under a social norm of in-group discrimination/out-group favoritism). *Fix $A(c) = \Theta \setminus \{(c, c)\}$ and consider a principal $p(t)$ of color $c_{p(t)} = c$. In equilibrium, if the principal selects an agent of type $\theta_{a(t),t} = (c, c)$, then $m(t) = \emptyset$. If the principal selects an agent of type $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, then*

1. *The principal will self-report: With self-reports, $m(t) = -c$.*
2. *Observers of the opposite color will not report: If $c_{o(t)} = -c$, then $m(t) = \emptyset$.*
3. *Observers of the same color will report: If $c_{o(t)} = c$, then $m(t) = -c$.*

Proof. Fix $A(c) = \Theta \setminus \{(c, c)\}$. Fix $A(-c)$ to one of the three norms defined in Proposition 2.3. By Lemma 2.1, if $\theta_{a(t),t} = (c, c)$, then $m(t) = \emptyset$. Assume for

¹⁰The existence of this new type of norm hails to the fact that, in the presence of competition, a principal can prevent her social color from reverting back to the default ($s_{p(t),t+1} = c$) if $\theta_{\mu(t),t} = (c, c)$. We provide a discussion of this result further below.

the rest of the proof that $\theta_{a(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$. In that case, either $m(t) = \emptyset$ or $m(t) = -c$. Immediate payoffs do not depend on the message. Exploiting the one-shot deviation principle, future payoffs depend on message $m(t)$ only with regard to payoffs in period $t + 1$ (see Lemma 2.3). If $m(t) = \emptyset$, then for all individuals of group c including $p(t)$, $\theta_{i,t+1} = (c, c)$. The set of individuals who satisfy $\theta_{i(t),t+1} \in \{(-c, -c), (-c, c), (c, -c)\}$ then includes only individuals of *physical* color $-c$. It follows from social norm $A(c) = \Theta \setminus \{(c, c)\}$ that in period $t + 1$, if $m(t) = \emptyset$, principals of group c only select agents of physical color $-c$. If, instead, $m(t) = -c$, then for all individuals of group c except $p(t)$, $\theta_{i,t+1} = (c, c)$, and for $p(t)$, $\theta_{p(t),t+1} = (c, -c)$. The set of individuals who satisfy $\theta_{i(t),t+1} \in \{(-c, -c), (-c, c), (c, -c)\}$ then includes all individuals of physical color $-c$ plus, as the unique exception to individuals of physical color c , individual $p(t)$. It follows that in period $t + 1$, principals of group c accept both, individuals of physical color $-c$ and individual $p(t)$ as agent. Payoffs in period $t + 1$ are affected by message $m(t)$ only if $c_{p(t+1)} = c$. If $c_{p(t+1)} = -c$, then by $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$, agent choice and thus, payoffs are unaffected by message $m(t)$.

Information disclosure thus depends on whether the sender of message $m(t)$ is better or worse off if individual $p(t)$ will be accepted as an agent by group c in period $t + 1$: If the sender is strictly better off, he sends $m(t) = -c$. If he is weakly worse off, he remains silent, $m(t) = \emptyset$. Clearly, individual $p(t)$ will self-report. Sending message $m(t) = -c$ yields her a higher probability to be selected as agent in period $t + 1$ and thus higher expected payoffs. This proves part 1 of the Lemma. Consider, next, an observer of the opposite color, $c_{o(t)} = -c$. The competitive environment implies that sending $m(t) = -c$ then strictly lowers the observer's expected payoffs in period $t + 1$: If he remains silent, $m(t) = \emptyset$, principals of group c will reject individual $p(t)$ as a match and select an alternative agent of physical color $-c$. With positive probability ($\frac{1}{n-c}$), the principal will select $o(t)$. Moreover, whenever another match of type (c, c) will be rejected, $m(t) = \emptyset$ yields the observer a higher probability to be selected as an alternative than message $m(t) = -c$.¹¹ It follows that if $c_{o(t)} = -c$, $m(t) = \emptyset$. This proves part 2 of the Lemma.

Consider, finally, an observer of the same physical color as the principal, $c_{o(t)} = c$. Sending $m(t) = -c$ then does not affect the probability of the observer to be selected as agent. Because $c_{o(t),t+1} = (c, c)$ with certainty, $o(t)$ will be rejected as agent by principals of group c with any message $m(t)$. Moreover, because

¹¹The conditional probability to be selected is $\frac{1}{|J^A(t+1)|} = \frac{1}{n-c}$ if $m(t) = \emptyset$ and $\frac{1}{|J^A(t+1)|} = \frac{1}{n-c+1}$ if $m(t) = -c$, respectively.

$(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$, the probability of $o(t)$ being selected as agent by principals of group $-c$ is also unaffected by $m(t)$. However, $m(t)$ affects the payoffs of $o(t)$ in the case that he is principal in period $t+1$: If $p(t+1) = o(t)$ and $\mu(t+1) = p(t)$ (probability $\frac{1}{n} \cdot \frac{1}{n-1}$), message $m(t) = -c$ will allow $o(t)$ to select his match as agent, earning him payoff πH . If $m(t) = -c$, on the other hand, he will have to reject his match and select an alternative agent, which only yields payoff $\pi L < \pi H$. It follows that if $c_{o(t)} = c$, $m(t) = -c$. This proves part 3 of the Lemma. \square

The next Lemma answers the question of when a principal of group c follows norm $A(c) = \Theta \setminus \{(c, c)\}$ given that after interacting with an agent of type $\theta_{\mu(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, this interaction yields message $m(t) = -c$ with probability $Prob[m = -c]$.

Lemma 2.7 (Agent choice under a social norm of in-group discrimination/out-group favoritism). *Fix $A(c) = \Theta \setminus \{(c, c)\}$ and consider a principal of color $c_{p(t)} = c$. If the match is of type $\theta_{\mu(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, then $a(t) = \mu(t)$. If the match is of type $\theta_{\mu(t),t} = (c, c)$, then the principal complies with the norm by selecting an alternative agent $j \neq \mu(t)$ of type $\theta_{j,t} \in \{(-c, -c), (-c, c), (c, -c)\}$ if*

$$\pi \left(1 - \frac{L}{H}\right) < Prob[m = -c] \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot \left(1 + \frac{n_c - 2}{n_c + 1} \cdot \frac{L}{H}\right) (1 - \pi),$$

(DC In-Group)

and selects the match, $a(t) = \mu(t)$, (does not comply) otherwise.

Proof. The proof is analogous to the proof of Lemma 2.5. Fix an equilibrium with $A(c) = \Theta \setminus \{(c, c)\}$ and consider the one-shot payoff-maximizing action of a principal $p(t)$ of color $c_{p(t)} = c$. If she selects her match, $a(t) = \mu(t)$, she yields immediate payoff πH . If she selects another agent, $a(t) = j \neq \mu(t)$, her immediate payoff is $\pi L < \pi H$. If she destroys the opportunity, immediate payoffs are zero. By Lemma 2.3, the continuation payoffs of the principal are affected by her choice of agent $a(t)$ only with regard to payoffs in period $t+1$. In particular, if her choice yields social color $s_{p(t),t+1} = c$, individual $p(t)$ will not be accepted as agent by group c in period $t+1$. If her choice yields social color $s_{p(t),t+1} = -c$, however, she will be accepted. It follows that $V_{p(t)}^{(c,-c)} > V_{p(t)}^{(c,c)}$. The difference in continuation payoffs calculates to

$$V_{p(t)}^{(c,-c)} - V_{p(t)}^{(c,c)} = \frac{n_c - 1}{n} \cdot \frac{1}{n-1} \cdot (1 - \pi)H + \frac{n_c - 1}{n} \cdot \frac{n_c - 2}{n-1} \cdot \frac{1}{n_c + 1} (1 - \pi)L.$$

The first term accounts for the expected payoffs from being accepted as match $\mu(t+1)$ by a principal of group c . The second term accounts for the expected payoffs from being selected as an alternative agent $a(t+1) \neq \mu(t+1)$ by a principal of group c . To calculate the second term, note that with probability $\frac{n_c-1}{n-1}$, another individual of group c is principal in period $t+1$. With conditional probability $\frac{n_c-2}{n-1}$, this principal is matched to another in-group member, but not $p(t)$. In that case, $\theta_{\mu(t+1),t+1} = (c, c)$. Social norm $A(c) = \Theta \setminus \{(c, c)\}$ implies that principal $p(t+1)$ will then reject his match and select an alternative agent of type $\theta_{j,t+1} \in \{(-c, -c), (-c, c), (c, -c)\}$. If $s_{p(t),t+1} = -c$, the set of acceptable agents includes all individuals of physical color $-c$ plus individual $p(t)$. The number of acceptable agents will then be $n_{-c} + 1$ implying that the probability that $p(t)$ will be selected is $\frac{1}{n_{-c}+1}$.

Equipped with $V_{p(t)}^{(c,-c)} - V_{p(t)}^{(c,c)}$, we are ready to determine the one-shot payoff-maximizing action $a(t)$ of principal $p(t)$ conditional on match $\mu(t)$. If the match conforms to the norm, that is, $\theta_{\mu(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, accepting the match yields expected continuation payoff $Prob[m = -c] \cdot V_{p(t)}^{(c,-c)} + (1 - Prob[m = -c]) \cdot V_{p(t)}^{(c,c)}$. No other action can yield a higher continuation payoff. In particular, selecting another agent $j \neq \mu(t)$ of type $\theta_{j,t} \in \{(-c, -c), (-c, c), (c, -c)\}$ yields identical expected continuation payoff, while selecting an agent j of type $\theta_{j,t} = (c, c)$ or destroying the opportunity yields weakly lower continuation payoff $V_{p(t)}^{(c,c)}$. It follows that if the match conforms to the norm, $\theta_{\mu(t),t} \in \{(-c, -c), (-c, c), (c, -c)\}$, the principal selects the match as agent. Consider, instead, the case that the match does not conform to the norm, $\theta_{\mu(t),t} = (c, c)$. Then the principal faces a trade-off between selecting the match and earning payoff $\pi H + \delta V_{p(t)}^{(c,c)}$, and selecting an alternative agent j of type $\theta_{j,t} \in \{(-c, -c), (-c, c), (c, -c)\}$ and earning payoff $\pi L + Prob[m = -c] \cdot \delta V_{p(t)}^{(c,-c)} + (1 - Prob[m = -c]) \cdot \delta V_{p(t)}^{(c,c)}$. The principal will comply with the norm and select an alternative agent if and only if

$$\begin{aligned} \pi(H - L) &< Prob[m = -c] \cdot \delta \cdot [V_{p(t)}^{(c,c)} - V_{p(t)}^{(c,-c)}] \\ \Leftrightarrow \pi \left(1 - \frac{L}{H}\right) &< Prob[m = -c] \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot \left(1 + \frac{n_c - 2}{n_{-c} + 1} \cdot \frac{L}{H}\right) (1 - \pi). \end{aligned}$$

Otherwise, the principal selects the match, $a(t) = \mu(t)$ (does not comply). □

From Lemmas 2.6 and 2.7 we can conclude:

Proposition 2.5 (In-group discrimination/out-group favoritism with endogenous information disclosure). *If interactions are competitive, a social norm of in-group discrimination/out-group favoritism $A(c) = \Theta \setminus \{(c, c)\}$ may exist. The norm can be sustained with observer-reports and self-reports. In particular, the norm exists for group $c \in \{\text{red}, \text{green}\}$ if and only if*

$$\pi \left(1 - \frac{L}{H}\right) < \text{Prob}[m = -c] \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot \left(1 + \frac{n_c - 2}{n_{-c} + 1} \cdot \frac{L}{H}\right) (1 - \pi),$$

(DC In-Group)

where $\text{Prob}[m = -c] = 1$ if information is disclosed through self-reports and $\text{Prob}[m = -c] = \frac{n_c - 1}{n - 1}$ if information is disclosed through observers.

Proof. With self-reports, $\pi^{-c} = 1$ (Lemma 2.6). By Lemma 2.7, the norm then exists if (ICC) holds with $\pi^{-c} = 1$. Consider Gossip. Then in any period t , $\text{Prob}[c_{o(t)} = c \mid c_{p(t)} = c] = \frac{n_c - 1}{n - 1}$. It then follows from Lemma 2.6 that $\pi^{-c} = \frac{n_c - 1}{n - 1}$. By Lemma 2.7, the norm then exists if (ICC) holds with $\pi^{-c} = \frac{n_c - 1}{n - 1}$. □

Note that competition for interactions plays an entirely different role in the existence result of in-group discrimination $A(c) = \Theta \setminus \{(c, c)\}$ (Proposition 2.5) than it does in the existence result of out-group discrimination $A(c) = \{(c, c)\}$ (Proposition 2.4). The crucial question regarding the existence of out-group discrimination concerns incentives for information disclosure. These exist only if the level of competition $\frac{L}{H}$ is sufficiently high. Under a social norm of in-group discrimination, incentives for information disclosure do not hinge on the level of competition. In fact, they exist even if the ratio $\frac{L}{H}$ goes to zero. Competition matters for a different reason: It allows the principal to *always* interact with an agent of type $\theta_{j,t} \in \{(-c, -c), (-c, c), (c, -c)\}$. This is decisive because in the absence of competition (see the benchmark model, section 2.2), principals lack the possibility to obtain social color $s_{p(t),t+1} = -c$ if they reject a match of type $\theta_{\mu(t),t} = (c, c)$. Because they cannot access the future benefits of being associated with the opposite color in that case, there are no incentives to follow a norm of in-group discrimination: $A(c) = \Theta \setminus \{(c, c)\}$ fails to exist. If, on the other hand, interactions are competitive, the principal can always obtain $s_{p(t),t+1} = -c$ (with positive probability) by simply selecting an agent who carries the opposite color in his type vector. This result is readily transferable to real-world situations: It is easier to credibly associate oneself

with the out-group and sustain a social norm of in-group discrimination in environments in which interactions with out-group members are more readily available.

2.4.5 Group Incentives for Discrimination

We have so far only considered individual incentives for discrimination. In our framework, individuals discriminate either because they face punishment in the form of ostracism from the in-group if they do *not* discriminate (in the case of a social norm of out-group discrimination, $A(c) = \{(c, c)\}$), or because they are favored as out-group members if they *do* discriminate (in the case of a social norm of in-group discrimination $A(c) = \Theta \setminus \{(c, c)\}$).

Discrimination is always detrimental when considering society as a whole. Taking the total sum of payoffs as a measure of welfare in society, the welfare-maximizing social norm is to be colorblind: If both groups behave colorblindly, total payoff generated each period is $H > 0$, of which each group on average receives a share equal to its population share $\frac{n_c}{n}$. If a group discriminates, it rejects the most productive agent in period t with positive probability, thereby destroying payoffs $H > 0$ (in the non-competitive benchmark model) or $(H - L) > 0$ (if interactions are assumed to be competitive), respectively. Competition can, however, give rise to group incentives for discrimination:

Proposition 2.6 (Group incentives for discrimination). *Fix any equilibrium, assuming that group $c \in \{\text{red, green}\}$ follows a colorblind social norm, $A(c) = \Theta$ (the norm always exists). Compared to this norm,*

(a) *A social norm of out-group discrimination, $A(c) = \{(c, c)\}$, yields strictly higher average payoff for any member of group c if and only if competition is sufficiently high, that is, if and only if*

$$\frac{L}{H} > \pi.$$

(b) *A social norm of in-group discrimination/out-group favoritism, $A(c) = \Theta \setminus \{(c, c)\}$, always yields strictly lower average payoff for any member of group c .*

Proof. The expected payoff of individual i , $c_i = c$, in period t is

$$E[u_i(t) | c_i = c] = \frac{n_c}{n} \cdot E[u_i(t) | c_{p(t)} = c] + \frac{n-c}{n} \cdot E[u_i(t) | c_{p(t)} = -c].$$

In equilibrium, payoff component $E[u_i(t) | c_{p(t)} = -c]$ is governed by the social norm of the opposite group, $A(-c) \in \{\Theta, \{(-c, -c)\}, \Theta \setminus \{(-c, -c)\}\}$. By $(c, c) \in A(-c) \Leftrightarrow (c, -c) \in A(-c)$, these payoffs do not depend on the variable component (social color) of the type of individual i and are thus independent of the social norm that group c follows. Social norm $A(c)$ affects only payoff component $E[u_i(t) | c_{p(t)} = c]$. Denote the average (across time) of this component \bar{u}^c . If group c follows a colorblind social norm, $A(c) = \Theta$, then

$$\bar{u}^c = \frac{1}{n_c} \cdot \pi H + \frac{n_c - 1}{n_c} \cdot \frac{1}{n - 1} (1 - \pi) H.$$

The first term accounts for the case that individual i is the principal: If $A(c) = \Theta$, she always selects her match as agent, earning payoff πH . The second term accounts for the case that another individual of group c is principal. In that case, individual i earns payoff $(1 - \pi)H$ if she happens to be the match of period t (probability $\frac{1}{n-1}$).

Assume that, instead, group c follows a social norm of out-group discrimination $A(c) = \{(c, c)\}$. Then

$$\bar{u}^c = \frac{1}{n_c} \cdot \left(\pi H - \frac{n-c}{n-1} \cdot \pi(H-L) \right) + \frac{n_c - 1}{n_c} \cdot \frac{1}{n-1} \left((1-\pi)H + \frac{n-c}{n_c-1} (1-\pi)L \right),$$

where the term $-\frac{n-c}{n-1} \cdot \pi(H-L)$ accounts for the loss in income on the side of the principal when having to reject an out-group match and the term $+\frac{n-c}{n_c-1} (1-\pi)L$ accounts for the added income on the side of in-group agents who will be selected when out-group agents are rejected. It is now easy to see that out-group discrimination *on average* yields a strictly higher payoff for individual i than a colorblind group norm if and only if $\pi(H-L) < (1-\pi)L \Leftrightarrow \frac{L}{H} > \pi$.

Consider, finally, the case that group c follows a social norm of in-group discrimination/out-group favoritism, $A(c) = \Theta \setminus \{(c, c)\}$. It is obvious that this norm cannot increase average payoffs among group c : Whenever the norm leads to a rejection of match $\mu(t)$, it destroys payoff πH for the principal and payoff $(1-\pi)H$ for the match, both of whom are members of group c . Even if the alternative agent who is selected in such a case were to belong to group c (an individual of type $(c, -c)$), the additional payoff to this individual would only amount to $(1-\pi)L$, which is strictly less than the cost incurred by the principal and the match. It follows that $A(c) = \Theta \setminus \{(c, c)\}$ always yields strictly lower average payoff for individuals of group c than a colorblind social norm. □

Given the multiplicity of social norms in certain parameter regions, a natural question to ask is whether there exists arguments for one social norm to be more likely to emerge than the other. Recall, from Proposition 2.3, that each group coordinates separately and individually on a social norm: There is no punishment or reward from out-group members regarding compliance.¹² Proposition 2.6 then answers the question of which social norm group c would *choose to coordinate on* if they had the possibility to ex-ante consult on the issue.¹³ Most intuitively, the group would never choose to in-group discriminate, $A(c) = \Theta \setminus \{(c, c)\}$. If $\frac{L}{H} \leq \pi$, all group members would ex-ante agree to coordinate on a colorblind norm, $A(c) = \Theta$. If $\frac{L}{H} > \pi$, however, all group members would ex-ante agree to coordinate on a social norm of *out-group* discrimination, $A(c) = \{(c, c)\}$.

Although discrimination is harmful for the society as a whole and must be enforced in individual interactions, if competition is sufficiently high, each group favors a discriminatory norm that restricts interactions to in-group members. This is irrespective of what type of social norm the opposite group follows. The situation is—from a group-level perspective—similar to a prisoners’ dilemma: If $\frac{L}{H} > \pi$, each group benefits from *unilaterally* deviating from a colorblind equilibrium. If both groups deviate, however, the resulting equilibrium with mutual discrimination generates lower average payoffs for any individual in society than the colorblind equilibrium did before. Nonetheless, out-group discrimination remains a group-level best response also in this situation. The finding is in line with literature in sociology that regards inter-group competition as a potential source of inefficient discrimination and in-group favoritism, see, for example Bobo and Hutchings (1996).¹⁴

Of course, a social norm of out-group discrimination can be enforced (i.e., exists) only if the conditions stated in Proposition 2.4 are satisfied. Note that the information disclosure constraint, $\frac{L}{H} > \frac{\pi}{1 + \frac{n-c}{n_c-1}(1-\pi)}$ is satisfied whenever the group prefers the norm, $\frac{L}{H} > \pi$. Whether the norm can be enforced then depends on whether ostracism from the in-group is sufficiently likely and painful to prevent individuals from interacting with the out-group, that is, whether the discrimination constraint (DC Out-Group) is satisfied. As norm violations are disclosed and punished only by in-group members, enforcement is generally easier for a majority group than for a mi-

¹²In particular, social norms $A(c)$ and $A(-c)$ are independent, see Proposition 2.3.

¹³For example, one could consider an extension of the game that includes an ex-ante one-shot public election among group members (say, in period $t = -1$) that decides on social norm $A(c)$.

¹⁴See also Bramoullé and Goyal (2016) for a microeconomic model that makes a similar claim.

nority group.¹⁵ Note that there exist parameter regions in which the group prefers a colorblind norm, $\frac{L}{H} \leq \pi$, but the information disclosure constraint (IC Out-Group) and discrimination constraint (DC Out-Group) are nonetheless satisfied. In these regions, out-group discrimination can emerge as an “unwanted” group norm, much in the spirit of the original “spontaneous discrimination” equilibria studied by Peski and Szentes (2013).

2.5 Conclusion

Discrimination can arise in tolerant societies via the coordination of groups on inefficient social norms that deliver reputational rewards to individuals who restrict their interactions to partners of a certain color. For such norms to be sustainable, information about the color of partners needs to be revealed to other members of the group. This essay shows that competition for interactions can generate incentives for information disclosure. In the presence of competition, discriminatory social norms yield benefits for one group at the expense of the other. Individuals disclose information about the color of partners in order to gain access to the (preferably small) group that benefits as well as to exclude others from it. Competition can also generate group-incentives for discrimination.

When interactions are competitive, both out-group and in-group discrimination can emerge as a social norm. While out-group discrimination requires third-party (observer) reports to be sustainable (Proposition 2.4, page 71), in-group discrimination can also be sustained with self-reports (Proposition 2.5, page 76). On one hand, this result speaks for in-group discrimination to be more likely to emerge than out-group discrimination. On the other hand, group-incentives for out-group discrimination (Proposition 2.6, page 77) point toward the opposite conclusion.

Our framework assumes that the default social color of individuals is their physical color. It also makes the assumption that matches are non-assortative. We conclude by discussing the consequences of relaxing those assumptions.

2.5.1 Empty Social Color

Following the framework of Peski and Szentes (2013), we have assumed that the default social color of individuals is their physical color, that is, $s_{i,t} = c_i$. It may

¹⁵See the right-hand side of (DC Out-Group), which strictly increases in $\frac{n_c-1}{n-1}$ (punishments). Moreover, $Prob[m = -c] = \frac{n_c-1}{n-1}$ (information disclosure).

seem that some of our results strongly hinge on this assumption. In fact, when information disclosure is endogenous, the assumption is surprisingly innocent. It is then equivalent to assuming that social color is empty by default, i.e., contains no information on group affiliation.

Consider the model with endogenous information disclosure and assume that additional to $s_{i,t} = c_i$ and $s_{i,t} = -c_i$, there exists an empty (= neutral) social color $s_{i,t} = \emptyset$. Let $s_{i,t} = \emptyset$ replace $s_{i,t} = c_i$ as the default: In period $t = 0$, $s_{i,t} = \emptyset$. If $i \neq p(t)$, then $s_{i,t+1} = \emptyset$. And if $i = p(t)$, then $s_{p(t),t+1} = m(t) \in \{\emptyset, c_i, -c_i\}$. We can then establish:

Proposition 2.7 (Empty social color). *Consider a model with endogenous information disclosure (Section 2.3 et sqq.), but assume that, by default, social color is empty, $s_{i,t} = \emptyset$ (see above). Then social norms in equilibrium are unchanged. In particular, Propositions 2.2–2.6 remain valid.*

Proof. Consider any period t , fixing $c_{p(t)} \in \{red, green\}$ as well as, in the case of observer-reports, $c_{o(t)} \in \{red, green\}$. Assume that the principal selects an agent, $a(t) \neq \emptyset$. Fix $c_{a(t)} \in \{red, green\}$. Note that type $(c_{a(t)}, c_{a(t)})$ and type $(c_{a(t)}, \emptyset)$ induce the same message space, namely $m(t) \in \{c_{a(t)}, \emptyset\}$. It follows that in equilibrium, they induce identical messages $m(t)$ and thus, assuming one-shot deviations, identical continuation payoff for the principal. This implies that for any colors $c_{a(t)} \in \{red, green\}$ and $c_{p(t)} \in \{red, green\}$, $(c_{a(t)}, c_{a(t)}) \in A(c_{p(t)}) \Leftrightarrow (c_{a(t)}, \emptyset) \in A(c_{p(t)})$. For this reason, in equilibrium, types (c_i, \emptyset) and (c_i, c_i) , for any $c_i \in \{red, green\}$, are equivalent: They behave identically given $A(c_i)$ (behavioral equivalence), they induce the same messages and actions by others (strategic equivalence), and yield the same payoffs (payoff equivalence). Without loss of generality, let $s_{i,t} = c_i$ whenever $s_{i,t} = \emptyset$. The result follows. □

Intuition derives from the fact that in a model with endogenous information disclosure, continuation payoffs and thus, social norms depend on the message space an individual induces when being selected as agent. By her physical color c_i an individual is already irrevocably associated with her in-group. If the individual is selected as agent, this allows for message $m(t) = c_i$ irrespective of her social color. Having a social color equal to one's physical color, $s_{i,t} = c_i$, does not affect the message space and is therefore equivalent with an empty social color $s_{i,t} = \emptyset$. The crucial question concerning an individual's social color is whether it relates her to the

opposite group, $s_{i,t} = -c_i$, and thus, allows for a message other than $m(t) = c_i$. This depends on the interactions the individual had in the past as well as on messages regarding those interactions, but not on whether, by default, $s_{i,t} = c_i$ or $s_{i,t} = \emptyset$.

2.5.2 Assortative Matching and Observation

We have assumed that the draw of match $\mu(t)$ —as well as in the case of endogenous information disclosure, the draw of observer $o(t)$ —is uniform random from the residual population. Given a principal of color $c_{p(t)} = c \in \{\text{red}, \text{green}\}$, the individual drawn as match $\mu(t)$ or observer $o(t)$ is of the same group with probability $\frac{n_c-1}{n-1}$ and of the opposite group with probability $\frac{n-c}{n-1}$. The probability to meet a person of a given group is thus equal to the share of that group in the residual population. This assumption is in line with the matching mechanism in Peski and Szentes (2013).

In some instances, however, it may be more realistic to assume that the likelihood to meet a member of a given group systematically deviates from these ratios. For instance, it may be that—for exogenous reasons such as neighborhood structure or group-level correlations of preferences and abilities—individuals $\mu(t)$ and $o(t)$ are disproportionately likely to be of the same physical color as the principal. In such a case the matching and observation processes would be assortative. How does assortativity affect the incentives for discrimination? As a concrete example, assume that a principal is more likely to be matched with (respectively, observed by) an individual in her spatial proximity. Is the propensity for discrimination then higher in a society with segregated neighborhoods or in a society with mixed neighborhoods (see Figure 2.5)?

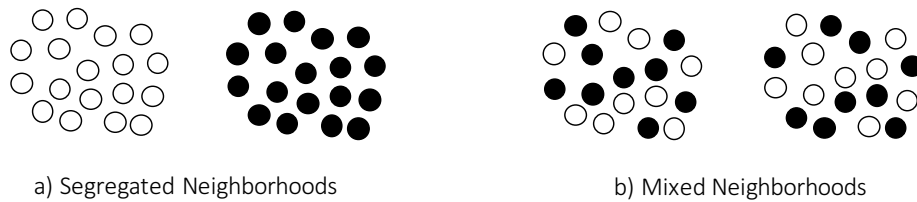


Figure 2.5: A population with two neighborhoods

For a formal analysis of assortative matching probabilities, let the probability that $c_{\mu(t)} = c_{p(t)} = c \in \{\text{red}, \text{green}\}$ be given by some constant $\rho^c \in (0, 1)$. If $\rho^c > \frac{n_c-1}{n-1}$, matching is *assortative*, while if $\rho^c < \frac{n_c-1}{n-1}$, it is *disassortative*. We continue to assume that conditional on matching with group $c' \in \{c, -c\}$, the probability to match with

any individual member of that group is uniform random. We can then observe:

Proposition 2.8 ((Dis-)assortative matching). *Assume that matching is (dis-)assortative with the conditional probability for $c_{\mu(t)} = c_{p(t)} = c$ being $\rho^c \in (0, 1)$. Then discrimination constraints in Propositions 2.1, 2.4 and 2.5 become*

$$\pi < .5 \cdot \delta \cdot \frac{\rho^c}{n} \cdot (1 - \pi) \quad (\text{DC Benchmark})$$

$$\pi \left(1 - \frac{L}{H}\right) < \text{Prob}[m = -c] \cdot \delta \cdot \frac{\rho^c}{n} \cdot \left(1 + \frac{1 - \rho^c}{\rho^c} \cdot \frac{L}{H}\right) (1 - \pi) \quad (\text{DC Out-Group})$$

$$\pi \left(1 - \frac{L}{H}\right) < \text{Prob}[m = -c] \cdot \delta \cdot \frac{\rho^c}{n} \cdot \left(1 + \frac{n_c - 2}{n_c + 1} \cdot \frac{L}{H}\right) (1 - \pi) \quad (\text{DC In-Group})$$

and the information disclosure constraint in Lemma 2.4 becomes

$$\frac{L}{H} > \frac{\pi}{1 + \frac{1 - \rho^c}{\rho^c} (1 - \pi)}. \quad (\text{IC Out-Group})$$

Other parts of Propositions 2.1–2.6 remain unaffected.

Proof. Omitted. (Incorporating ρ^c in the proofs of Propositions 2.1–2.6 yields the result.) □

Under the assumption that information disclosure relies on observer-reports, assortativity also affects $\text{Prob}[m = -c]$ in constraints (DC Out-Group) and (DC In-Group). Recall that only in-group members have an incentive to report. Applying the the same assumptions to the draw of $o(t)$ as to the draw of $\mu(t)$ then implies $\text{Prob}[m = -c] = \rho^c$.¹⁶

We can conclude: Assortativity does not qualitatively alter the main results of the model. It does, however, make discriminatory norms easier to enforce. Once we acknowledge that norms are coordinated on and enforced separately within each group $c \in \{\text{red}, \text{green}\}$,¹⁷ this result is intuitive: Assortativity implies a higher probability of interaction and observability within the group and is thus associated with a higher reputational cost of deviation. In a model of neighborhood assortativity (Figure 2.5), spatial segregation would be associated with a higher propensity for norm compliance and thus, a higher propensity for discrimination.

¹⁶To be specific, in the case of out-group discrimination, $\text{Prob}[m = -c] = \rho^c$ if (IC Out-Group) is satisfied and $\text{Prob}[m = -c] = 0$ otherwise.

¹⁷See Propositions 2.1 and 2.3: The social norm of group c is independent of the norm followed by the other group, $A(-c)$.

References

- Akerlof, George A., and Rachel E. Kranton.** 2000. "Economics and Identity." *The Quarterly Journal of Economics*, 115(3): 715–753.
- Ali, S. Nageeb, and David A. Miller.** 2016. "Ostracism and Forgiveness." *American Economic Review*, 106(8): 2329–2348.
- Arrow, Kenneth J.** 1973. "The Theory of Discrimination." In *Discrimination in Labor Markets.*, ed. Orley C. Ashenfelter and Albert Everett Rees, 3–33. Princeton: Princeton University Press.
- Becker, Gary S.** 1957. *The Economics of Discrimination*. Chicago: The University of Chicago Press.
- Bénabou, Roland, and Jean Tirole.** 2012. "Laws and Norms." *IZA Discussion Paper No. 6290*.
- Bobo, Lawrence, and Vincent Hutchings.** 1996. "Perceptions of Racial Group Competition: Extending Blumer's Theory of Group Position to a Multiracial Social Context." *American Sociological Review*, 61(6): 951–972.
- Bramoullé, Yann, and Sanjeev Goyal.** 2016. "Favoritism." *Journal of Development Economics*, 122: 16–27.
- Choy, James P.** 2017. "Social Division with Endogenous Hierarchy." *The Economic Journal*, Forthcoming.
- Coate, Stephen, and Glenn C. Loury.** 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review*, 83(5): 1220–1240.
- Dal Bó, Pedro.** 2007. "Social Norms, Cooperation and Inequality." *Economic Theory*, 30(1): 89–105.
- Darity, William A. Jr., and Rhonda M. Williams.** 1985. "Peddlers Forever?: Culture, Competition, and Discrimination." *The American Economic Review*, 75(2): 256–261.
- Eeckhout, Jan.** 2006. "Minorities and Endogeneous Segregation." *Review of Economic Studies*, 73: 31–53.

- Eguia, Jon X.** 2015. “Discrimination and Assimilation.” mimeo.
- Holzer, Harry J., and Keith R. Ihlanfeldt.** 1998. “Customer Discrimination and Employment Outcomes for Minority Workers.” *The Quarterly Journal of Economics*, 113(3): 835–867.
- Kandori, Michihiro.** 1992. “Social Norms and Community Enforcement.” *Review of Economic Studies*, 59: 63–80.
- Lang, Kevin, and Jee-Yeon K. Lehmann.** 2012. “Racial Discrimination in the Labor Market: Theory and Empirics.” *Journal of Economic Literature*, 50(4): 959–1006.
- Mailath, George J., and Larry Samuelson.** 2006. *Repeated Games and Reputations*. New York:Oxford University Press.
- Mailath, George J., Larry Samuelson, and Avner Shaked.** 2000. “Endogenous Inequality in Integrated Labor Markets with Two-Sided Search.” *American Economic Review*, 90(1): 46–72.
- Massey, D., and N. Denton.** 1993. *American Apartheid*. Cambridge, Mass.:Harvard University Press.
- McAdams, Richard H.** 1995. “Cooperation and Conflict: The Economics of Group Status Production and Race Discrimination.” *Harvard Law Review*, 108(5): 1003–1084.
- Peski, Marcin, and Balasz Szentes.** 2013. “Spontaneous Discrimination.” *American Economic Review*, 103(6): 2412–2436.
- Phelps, Edmund S.** 1972. “The Statistical Theory of Racism and Sexism.” *American Economic Review*, 62(4): 659–661.
- Ramachandran, Rajesh, and Christopher Rauh.** 2016. “Discrimination Without Taste - How Discrimination Can Spillover and Persist.” mimeo.

Chapter 3

Corrupted Votes and Rule Compliance

Authors: Arno Apffelstaedt and Jana Freundt

Abstract: We study—using an online experiment with international subjects—how compliance with elected rules of conduct is affected by having experienced an election in which (1) subjects are asked for money to make their vote count, (2) subjects are offered money for voting differently, or (3) subjects with low household income are excluded from the ballot. We find strong and significant reductions in compliance rates across the population after such “corrupt elections”, but only if elected rules ask subjects to behave prosocially. Treatment effects seem to be driven by intrinsic concerns about procedural aspects of the election mechanism and are prevalent mainly among individuals who—in a questionnaire that is presented as an unrelated survey two weeks after the experiment—express high value for democratic institutions and low value for bribing and (political) lobbying in the real world.

Keywords: Endogenous Institutions, Corruption, Rule Compliance

JEL Codes: D72, D91, B55, C92

3.1 Introduction

An influential stream of papers in public and political economics suggests that democratic institutions may affect behavior.¹ Frey (1997), for example, finds that tax

¹There is a related literature in organizational economics that studies the value of “democratic” decision making mechanism within firms. Bonin, Jones and Putterman (1993), Black and Lynch (2001) and Zwick (2004), for example, provide empirical support that employee participation is associated with increased worker productivity. On a general account, Bartling, Fehr and Herz (2014) are able to demonstrate experimentally that many people yield intrinsic value from decision rights.

compliance is higher in Swiss cantons that see more democratic participation. Bardhan (2000) shows that South Indian farmers are more likely to follow irrigation rules if they partake in crafting them. Experimentally, Tyran and Feld (2006), Ertan, Page and Putterman (2009) and Sutter, Haigner and Kocher (2010), among others, demonstrate that punishments and rewards have greater impact on contributions to a public good when they are implemented by majority vote rather than exogenously by a computer. Dal Bó, Foster and Putterman (2010) provide experimental evidence of a similar ‘democracy effect’ in co-ordination games.² A conclusion that can be drawn from this literature is that giving citizens decision rights through elections and referenda can bring important efficiency gains to societies.

In many countries, however, promises of “free and fair” elections are undermined by practices ranging from systematic vote buying to arguably unintentional disenfranchisement of poor voters.³ Similar to how the introduction of a democratic procedure can generate positive behavioral responses, perceived malpractice and voter manipulation during elections may lead to negative behavioral consequences. In this essay, we test this hypothesis using a novel online experiment. The experiment studies how vote buying and voter disenfranchisement during a referendum affects the willingness of individuals to comply with elected rules asking them to behave pro-socially (to redistribute income) and with elected rules asking them to behave selfishly (to not redistribute). To our knowledge, this is the first experimental study on whether the well-documented positive behavioral effects of democratic institutions are sensitive to electoral malpractice. In comparison to earlier studies on ‘democracy effects’, our experiment allows us to say more about the psychological mechanisms underlying behavior and treatment effects. We establish a strong negative (intrinsic) effect of electoral malpractices on compliance with pro-social rules: When votes have been bought or parts of the electorate been excluded from the ballot, subjects comply sig-

²This list of studies is not meant to be exhaustive. See, e.g., Dal Bó (2014) for further studies.

³In a survey study in Argentina from 2002, for example, 35% of respondents reported to have observed the distribution of gifts by political parties in their neighborhoods during election campaigns and 12% of low-income respondents reported to have received something from a political party or candidate (Brusco, Nazareno and Stokes, 2004, pp. 69-70). According to a list experiment by Gonzalez-Ocantos et al. (2012) (a technique that usually assures to minimize social desirability biases in sensitive survey questions) more than 24% of registered voters reported to have been offered some sort of gift for their vote after the 2008 Nicaraguan municipal election. Examples for arguably unintentional voter disenfranchisement are restrictive ID laws (De Alth, 2009) or felon disenfranchisement (Manza and Uggen, 2008) in some states of the US. In 2017 alone, allegations of voter fraud have led to violent demonstrations in Turkey, Venezuela, Indonesia and the US, among other countries. A systematic, world-wide analysis of electoral malpractices and survey-based evidence of voters’ expressed dissatisfaction with biased electoral procedures can be found, for instance, in Norris (2014).

nificantly less with elected rules that ask them to redistribute. Maybe surprisingly, we find no significant treatment effects on compliance with selfish rules.

We study redistribution choices in experimental societies made up of 100 individual subjects. Subjects are recruited online via the platform Prolific.ac.⁴ The experiment revolves around the decision of whether one should redistribute income earned through luck to another member of the society who was unlucky (i.e., did not receive any income). We implement this decision with a binary one-shot dictator game: Each subject in the society has to decide conditional on receiving income whether she wants to $Give_i \in \{0, 1\}$ thirty percent of her income to a randomly matched person $j \neq i$ who did not receive income. Before subjects decide whether to redistribute, there is a referendum on the right “code of conduct.” Each subject can vote for a (society-wide) code that promotes giving (*Rule:Give*) or for a code that promotes non-giving (*Rule:Don't*). After the referendum, subjects decide (individually and anonymously) whether they want to $Give_i | Rule:Give \in \{0, 1\}$ conditional on *Rule:Give* being elected and whether they want to $Give_i | Rule:Don't \in \{0, 1\}$ conditional on *Rule:Don't* being elected. We are interested in how voluntary compliance with each of the two rules depends on electoral malpractice (in the form of vote buying or partial disenfranchisement) being present during the referendum.⁵ The hypothesis guiding our analysis is that compliance with both rules should be *lower* in societies that experience malpractice during the referendum compared to the levels of compliance observed in a society that did not experience electoral malpractice.

Using different treatment groups (each consisting of a society with 100 subjects), we introduce interventions to the referendum that may either lead to some voters being excluded from the ballot (= partial disenfranchisement) or to some votes not being representative of the true opinion of their issuer (= vote buying). Our interventions are the introduction of a small voting fee (the votes of subjects who do not pay are not counted towards the referendum), monetary offers to all subjects if they vote for the rule opposite to their first choice (vote buying), and an exclusion of all

⁴Prolific.ac has a subject pool of about 40.000 people and administers recruiting and payment. The Prolific.ac subject pool consists of individuals out of whom 60% are male, 26% are students, 85% speak English as a first language, roughly 60% have the UK nationality and 25% the US nationality. The remaining subjects have all kinds of different nationalities. The median age is 27. Education levels vary from no formal education (3%), college education (41%), undergraduate (33%) or graduate (18%) education to doctoral degrees (4%). See <https://www.prolific.ac/demographics> (accessed November 11th, 2017).

⁵Complying with the elected code of conduct is entirely voluntary: There is no formal punishment involved with deviation. There is also no possibility for other subjects to punish the choice of individual i .

subjects from the ballot whose household income is below a certain threshold (GBP 40,000). A baseline treatment in which the votes of all 100 subjects are counted in an unbiased way serves as the comparison.

We choose to study behavior in one-shot dictator games primarily for two reasons. The first reason is that non-binding rules in this domain should mainly work by their normative appeal. In particular, (classical) co-ordination issues as well as punishment concerns that exist in other games should not play a role in this setting.⁶ This makes dictator games particularly well suited for the analysis of whether procedural changes in how an election is conducted affect the intrinsic motivation of subjects to follow rules.⁷ For reasons we discuss in the next paragraph we hypothesize that rules should have higher normative appeal when they were selected in an inclusive and unbiased way, that is, with a referendum that did not involve vote buying or disenfranchisement. The second reason is that we aim to create an experimental situation in which people disagree about the “right” code of conduct and hence, potentially, vote for different rules. Note, importantly, that there is no efficiency-dominant rule. *Rule:Give* and *Rule:Don’t* differ only in their distributive nature. Earlier studies have shown that people differ in their judgements regarding whether income received through luck should be redistributed, see, in particular, Cappelen et al. (2007) and Almås, Cappelen and Tungodden (2017). Our setup allows us to study behavior under rules that promote “egalitarian” values (*Rule:Give*) and rules that promote “libertarian” values (*Rule:Don’t*).⁸

Finer details of our experimental design are meant to identify the psychological determinants of behavior that underlie rule-compliance and treatment effects. Research in psychology and behavioral economics suggests that procedural aspects of

⁶Earlier experiments on the behavioral effects of democratic elections have primarily looked at repeated public good games, trust games, and co-ordination games, see e.g., Tyran and Feld (2006) and Dal Bó, Foster and Putterman (2010). In those games, expectations about the behavior of other subjects are likely to play a more important role than they do in a dictator game. While there are no classical co-ordination incentives in one-shot dictator games—conditional on being a dictator, the strategies of other agents cannot influence a subject’s monetary payoff—there might be “psychological” co-ordination incentives arising from the wish to align one’s behavior with what others do or value. Our experiment is designed to test for such incentives, see the next paragraph.

⁷Dictator games have been chosen in earlier studies for similar reasons, see, for example, Krupka and Weber (2013), albeit not to our knowledge in studies on the effects of democracy on behavior. Note also that dictator games, in comparison to other interesting games in which rule-compliance is key—for example, games used to study cheating or lying behavior (Fischbacher and Föllmi-Heusi, 2013; Gächter and Schulz, 2016)—, do not entail the possibility that with non-compliance a subject can punish the *experimenter* for a procedure she perceives as unfair.

⁸Our use of the words “egalitarian” and “libertarian” follows Almås, Cappelen and Tungodden (2017).

decision making can affect preferences directly. In particular, people seem to care about the “fairness” of decision making processes (see, e.g., Tyler, 1990; Frey, Benz and Stutzer, 2004; Cappelen et al., 2013) as well as about personally partaking in them (see, e.g., Bonin, Jones and Putterman, 1993; Bardhan, 2000; Bartling, Fehr and Herz, 2014). Vote buying and partial disenfranchisement during elections is certain to affect preferences on the latter domain. Intuitively, preferences concerning the fairness of the decision making process should also be affected. The view that procedural concerns may lower the normative appeal of elected rules and thus, directly affect the willingness of people to comply is related to theories of “legitimate authority” (Weber, 1978; Tyler, 2006; Akerlof, 2017). We control for three aspects that might affect a subject’s decision to comply with rules in the dictator game apart from such concerns: (1) her preferences regarding the “right” code of conduct, (2) her behavior in the absence of a rule, and (3) her beliefs about the behavior of other subjects. To control for (1), we introduce our treatment interventions only after all subjects have stated a preference for the rule (*Rule:Give* or *Rule:Don’t*) they want to vote for. This allows us to control for the unbiased vote of a subject in all treatments—even if this vote might not count towards the final referendum.⁹ We control for (2) by introducing a prior round of the dictator game to our experiment in which subjects decide whether to $Give_i \in \{0, 1\}$ without knowing that there will be a second round that includes the election of a code of conduct. This allows us to identify a subject as a “natural” giver or non-giver—a categorization that turns out to play an important role in our analysis. Instead of giving subjects information about the behavior of other participants in this round—which might induce undesired punishment behavior in the second round of the dictator game following the referendum—, we present them with partial information about redistribution choices in our experimental pilot. By varying this information randomly on a subject-by-subject basis, we generate exogenous variance in the beliefs about the behavior of other subjects. This helps us to *causally* identify (3): The role of others in guiding behavior.¹⁰ Beliefs about the voting and compliance behavior of other subjects as well as beliefs about the impact of manipulative interventions on the referendum outcome are elicited (in an incentive compatible way) from every subject at the end of the experiment. Our main finding regarding the psychological determinants of

⁹This control follows the identification procedure introduced by Dal Bó, Foster and Putterman (2010).

¹⁰For example, we can use variance in the information we give subjects after round 1 of the dictator game to instrument for variance in beliefs about the behavior of other subjects in round 2.

behavior is that beliefs about the behavior of other subjects seem to play little to no role in explaining our treatment effects. Rather, subjects seem to react intrinsically to violations of the democratic ideal that elections should be inclusive and unbiased.

We complement our experiment with an extensive questionnaire on subjects' standpoints regarding various political issues such as redistribution, corruption, democratic values, and personal trust in institutions. To prevent the risk of spillovers from exposure to different treatments to questionnaire answers, the questionnaire is presented as an unrelated survey (using a different design and researcher profile) and is sent to the same people about two weeks after they participated in the experiment. We use the questionnaire to study whether self-reported standpoints on the value of democratic institutions correlate with reactions to electoral manipulation in the experiment. Indeed, we find that our treatment effects are mainly driven by subjects who self-report to have a high valuation for democratic institutions.

Indicative evidence for the hypothesis that electoral malpractice affects the willingness of people to comply with social rules and laws can also be found in observational data. In answers gathered from the World Values Survey (see Figure 3.1) the level of electoral malpractice perceived in a country is positively correlated with individual judgments regarding the justifiability of breaking rules, ranging from wrongfully claiming government benefits to cheating on taxes. However, because the level of malpractice is difficult to randomize in real elections, causality is hard to establish in the field. Where this is possible, researchers then generally have to rely on surveys to measure aggregate effects on behavior.¹¹ Individual level behavioral measures of voluntary rule-compliance are almost impossible to come by due to the difficulty to control for formal and informal deterrence measures that are in place in the field. An additional comparative advantage to using real world data is that our experimental framework enables us to study the psychological mechanisms driving treatment effects.

By relying on direct instead of indirect behavioral measures of support and dissatisfaction among citizens, political scientists have mostly taken a different approach towards assessing people's acceptance of elected institutions. Extensive survey stud-

¹¹For example, Berman et al. (2014) sent letters to a random sample of Afghan polling stations announcing that researchers would photograph election results and that these photographs would later be compared to certified results. This threat of control seems to have reduced election fraud (see also Callen and Long, 2015). The authors rely on a post-election survey to measure the effect of this treatment on attitudes towards government, of which "the willingness to report insurgent behavior to security forces" is the measure closest to what we are after. They find that sending a letter increases this willingness by 2.5 to 3 percentage points, which is statistically significant and supports our hypothesis.

How justifiable is...?

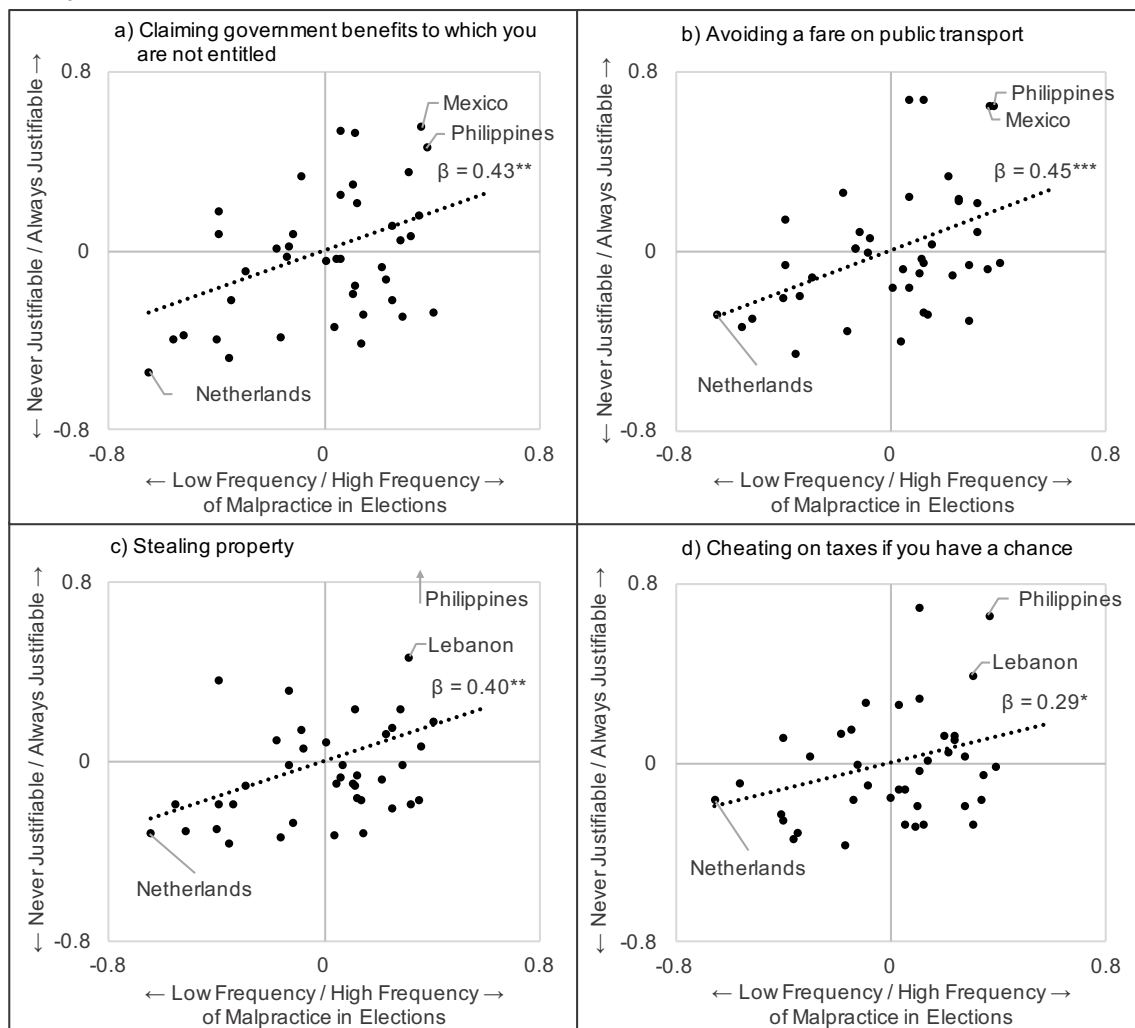


Figure 3.1: Country-level correlations between citizens' perceived frequency of malpractice in elections and their statements about the justifiability of violating rules and laws. Source: Country averages calculated from the WVS (2014). The figures plot the average answers in a country to questions V198-V201 against an index of perceived malpractice in elections. This index is calculated from the average of answers in a country to questions V228 B,C,D,G, and H (How often do the following things occur in your country? B: Opposition candidates are prevented from running, C: TV news favor the governing party, D: Voters are bribed, G: Rich people buy elections, H: Voters are threatened with violence at the polls). We have normalized the data to show relative deviations from the average across all countries. For example, in panel d), Lebanon's data point is (0.30, 0.38) meaning that it has a 30% higher measure of perceived malpractice and 38% higher measure of justifiability for tax cheating than the average country in our sample. The β -coefficients are from univariate OLS regressions without intercept: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$ assuming OLS standard errors.

ies of whether electoral malpractice undermines citizens’ expressed support for institutions is provided by Norris (2014). An experimental approach to eliciting such direct support is taken by, for example, Dickson, Gordon and Huber (2015), who measure the legitimacy of an institution by observing whether participants help or hinder an authority in punishing free-riders in a public good game. We are not aware of an experimental study that is trying to test what we are after.

The remainder of the chapter is structured as follows. Section 3.2 presents the experimental design in detail together with the predictions and identification strategy. Sections 3.3 and 3.4 present our results: We first estimate the average effect of vote manipulation on compliance rates and then study determinants of individual rule compliance. We conclude in section 3.5. Experimental instructions, screenshots, and the questionnaire can be found in the appendix to this chapter.

3.2 Experimental Design

The design of our online experiment is based on a referendum among 100 subjects on the preferred “code of conduct” regarding behavior in a dictator game. For each treatment, 100 subjects participate in a lottery that has one of them winning GBP 100. They are informed that the computer will unequally distribute lottery tickets among the 100 participants: 50 subjects will be “receivers” who get 10 lottery tickets each, while the remaining 50 subjects will be “non-receivers” and get no tickets. One of the 500 distributed lottery tickets is the winning ticket. We use this setup to construct a dictator game with role uncertainty: Before learning whether one is a receiver or a non-receiver of tickets, each subject is asked to (privately) decide whether—in case of being a receiver—she wants to give three out of ten lottery tickets to a randomly selected non-receiver.¹² In other words, each subject decides whether she wants to redistribute chances to win that she received through luck to another participant who was unlucky. In each session, we implement two rounds of this dictator game. Round 1 is a simple individual decision, the choice of individual i in this part is coded $Give_i \in \{0, 1\}$. In round 2, before subjects play the the dictator game again, they hold a referendum on a “code of conduct” for the whole group of 100 subjects. All subjects vote (privately) for either *Rule:Give*

¹²Subjects are informed that in the case of being a receiver (50% probability), their decision is automatically implemented and determines the number of lottery tickets for them and for one random other. They are also informed that in the case of being a non-receiver (50% probability), their decision does not play a role for the distribution of lottery tickets.

(“everybody should give”), $Vote_i = 1$, or for *Rule:Don’t* (“everybody should not give”), $Vote_i = 0$. After the referendum, each individual decides privately whether she wants to $Give_i|Rule:Give \in \{0, 1\}$ conditional on *Rule:Give* being elected and whether she wants to $Give_i|Rule:Don’t \in \{0, 1\}$ conditional on *Rule:Don’t* being elected. There is no (monetary) punishment involved in not following the elected rule.

Treatments differ in whether or not there is malpractice during the referendum and, if there is malpractice, in the form of malpractice introduced. We introduce treatment interventions *after* subjects have voted, but *before* they take decisions $Give_i|Rule:Give$ and $Give_i|Rule:Don’t$. The baseline treatment (*T_Baseline*) implements a simple majority vote. After voting, subjects are informed that “the rule that receives more votes in total will be implemented as the code of conduct.” The other three treatments allow for the possibility that either, some votes are not counted towards the majority vote, or that the final votes may have been manipulated. In *T_Pay4Vote*, after voting, subjects see a screen that asks them to pay GBP 0.20 to make their vote count and informs them that the code of conduct will be selected by majority vote among those subjects who accepted to pay. In *T_Bribe*, subjects see a screen that offers them a bonus payment of GBP 0.20 if they reverse their vote and informs them that the code of conduct will be selected by majority vote after each subject has decided to either accept or reject this offer. Finally, in *T_ExcludePoor*, subjects are informed that the code of conduct will be selected by majority vote among subjects with an annual household income above GBP 40,000. They are also informed whether this means that their personal vote is counted or not.¹³ The prediction guiding our analysis is:

Prediction 3.1 (Malpractice Effect). *The manipulation of electoral processes lowers voluntary compliance with elected rules:*

- (a) $E(Give_i|Rule:Give, Malpractice = 1) - E(Give_i|Rule:Give, Malpractice = 0) < 0$
- (b) $E(Give_i|Rule:Don’t, Malpractice = 1) - E(Give_i|Rule:Don’t, Malpractice = 0) > 0$

*In our experiment, $Malpractice = 1$ if individual i is in treatment *T_Pay4Vote*, *T_Bribe*, or *T_ExcludePoor*, and $Malpractice = 0$ if individual i is in treatment *T_Baseline*.*

¹³To identify a subject as having a household income above or below GBP 40,000, we use self-declared information provided to us (with consent of the participants) by the online-platform *Prolific.ac*.

3.2.1 Theoretical Framework

To fix ideas, consider the following simple theoretical framework.¹⁴ Consider, first, the decision to give in the absence of a code of conduct. Let $u_i(\text{Give}_i)$, $\text{Give}_i \in \{0, 1\}$ denote the utility of individual i when deciding to give or not give, respectively. Individual i then chooses to give if and only if

$$\Delta u_i(\text{Give}) := u_i(1) - u_i(0) \geq 0.$$

Classical economic theory would predict that $\Delta u_i(\text{Give})$ is negative. A positive $\Delta u_i(\text{Give})$ may reflect social preferences of individual i or “warm glow”.¹⁵ People might also want to align their behavior with anticipated giving behavior of others, driven by preferences for conformity (Bernheim, 1994; Bénabou and Tirole, 2012) or positive reciprocity (Fehr and Gächter, 2000). We will call those who give *Givers* and those do not give *Non-Givers* throughout the analysis. Let $\Delta u_i(\text{Give})$ be distributed in the population with cumulative density function $F[\cdot]$. In the absence of a rule, the share of *Givers* in the population is then given by $1 - F[0]$ as illustrated in Figure 3.2, panel a), below.

Now consider the case in which there exists a democratically elected code of conduct that either promotes giving, *Rule:Give*, or promotes non-giving, *Rule:Don't*. Theories of “legitimate authority” (e.g., Weber, 1978; Tyler, 2006; Akerlof, 2017) suggest that if a rule has come into force by a fair procedure, “people feel that they ought to defer [its] decisions and rules, following them voluntarily out of obligation rather than out of fear of punishment or anticipation of reward.” (Tyler, 2006, p.375). This is in line with earlier literature in psychology and behavioral economics which suggests that procedural aspects of decision making affect preferences directly (Tyler, 1990; Frey, Benz and Stutzer, 2004; Cappelen et al., 2013; Bartling, Fehr and Herz, 2014, among others). If people care to align their behavior with others, elected rules might change behavior because they provide a signal about what others do and value (Basu, 2015; Akerlof, 2016). Earlier experiments (e.g., Tyran and Feld, 2006;

¹⁴We provide a framework regarding the effect of our treatments on giving behavior. We extend this framework to cover voting behavior in the appendix.

¹⁵Typical examples in standard dictator games would be Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Andreoni (1989, 1990). Note however that due to individual i distributing lottery tickets, these theories can explain positive giving rates in our setting only if endowments are understood in an *ex ante* sense, that is, under the assumption that individual i has preferences over the distribution of winning probabilities. Saito (2013), for example, offers a model that introduces such preferences.

Sutter, Haigner and Kocher, 2010; Dal Bó, Foster and Putterman, 2010) confirm that endogenously elected institutions have the power to change behavior, but do not disentangle the psychological reasons why. Our experiment is designed to provide more insights into the psychological mechanism. For the theoretical framework, we shall take a “reduced form” approach: Assume that complying with a democratically elected rule adds fixed utility $\bar{u}^B \geq 0$ to $u_i(0)$ or $u_i(1)$, respectively. It then follows that individual i chooses to give iff

$$\Delta u_i(\text{Give}) \geq \begin{cases} -\bar{u}^B & \text{under } \textit{Rule:Give}, \\ +\bar{u}^B & \text{under } \textit{Rule:Don't}. \end{cases}$$

Compared to the case without a code, the share of givers in the population increases or decreases, see Figure 3.2, panels b) and c). Note that the rule should only affect behavior of those individuals who in the absence of a code would have chosen the opposite action. For instance, a democratically elected code that promotes giving (*Rule:Give*) may convince a *Non-Giver* to give, but will leave the behavior of a “natural” *Giver* unaffected.

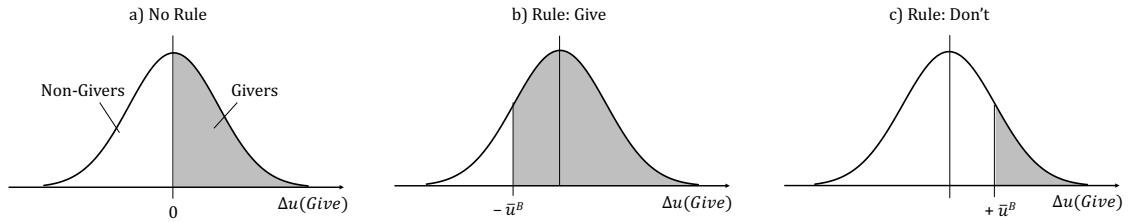


Figure 3.2: Theory: Share of Givers with and without rules

How is rule compliance affected by attempts to disenfranchise or manipulate voters during the election of a code? Again, we take a simple reduced form approach and assume that our interventions lower the utility to follow the elected rule by a constant $\Delta\bar{u}^M > 0$. This is line with both theoretical explanations laid out above: When the elected code does not represent the true preferences of all voters, this might affect the intrinsic motivation of a subject to follow the rule. It will also introduce noise into the signaling process of underlying values. In both cases, malpractice lowers the incentives to follow a given code: Individual i chooses to give iff

$$\Delta u_i(\text{Give}) \geq \begin{cases} -(\bar{u}^B - \Delta\bar{u}^M) & \text{under } \textit{Rule:Give}, \\ +(\bar{u}^B - \Delta\bar{u}^M) & \text{under } \textit{Rule:Don't}. \end{cases}$$

First and foremost, manipulating or disenfranchising voters thus leads people to revert back to their individually preferred behavior: As $\Delta\bar{u}^M$ increases, a lower share of *Non-Givers* will follow *Rule:Give*, see Figure 3.3, panel b). Similarly, a lower share of *Givers* will be willing to follow *Rule:Don't*, Figure 3.3, panel c). As $\Delta\bar{u}^M$ becomes sufficiently large such that $\bar{u}^M - \Delta\bar{u}^M$ turns negative, people may even turn against rules that match their “natural” giving preferences. For example, it is theoretically possible that giving under *Rule:Give* will deteriorate below rates observed in the absence of a code, although such a strong reaction might be unlikely to be observed in the experiment.

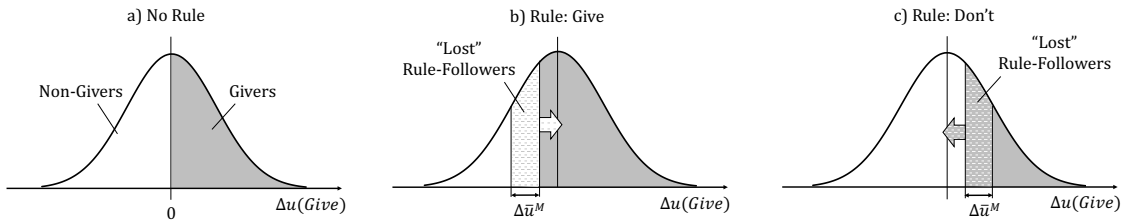


Figure 3.3: Theory: Effects of Interventions (Electoral Malpractice) on Rule-Compliance

3.2.2 Experimental Procedures

We will now detail all steps of an experimental session. For each treatment, 100 individual subjects are recruited on the online platform Prolific.ac with a small, fixed base payment and the prospect that one of 100 participants will win GBP 100. Before a participant starts the experiment, she receives detailed instructions on how the lottery tickets will be distributed (see Appendix D). Control questions at the end of each screen have to be answered correctly in order to proceed with the experiment.¹⁶ Participants are informed that there are two rounds but they only learn about the referendum that will take place in round 2 after having completed round 1. One round is randomly drawn to determine the final distribution of lottery tickets. All decisions are taken anonymously.

Timeline of Experimental Session. In round 1, each subject plays the dictator game ($Give_i \in \{0, 1\}$) individually. After the decision, subjects do not receive feedback about the giving behavior in their cohort. Instead, we show each subject

¹⁶We observe the number of times an individual tried to proceed without having answered all questions correctly. The number of such mistakes is generally small and has no explanatory power for our results.

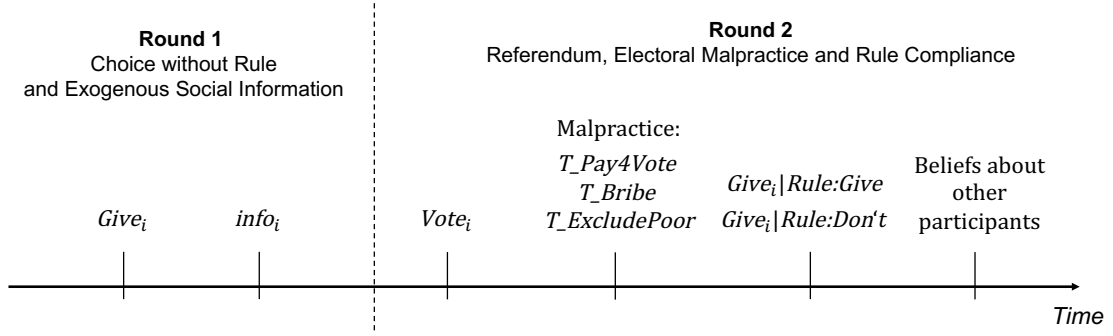


Figure 3.4: Timeline of Experimental Session

exogenous information on the giving decisions of five participants from an earlier session. An independent random draw determines if a subjects sees a sample where two out of five participants chose to give ($info_i = 2$) or one where four out of five participants chose to give ($info_i = 4$).

Participants then move to round 2, where they are informed that in this round, there will be a code of conduct for behavior in the dictator game. Every subject votes ($Vote_i \in \{0, 1\}$) on whether she prefers to have a code of conduct for all 100 subjects that says “give” ($Rule:Give$) or one that says “don’t give” ($Rule:Don't$). Treatments vary between subjects and are introduced after the vote. In $T_Pay4Vote$, each participant now decides whether she wants to pay GBP 0.20 to make her vote count. In T_Bribe , each participant decides whether she wants to accept GBP 0.20 and reverse her original vote. In $T_Baseline$ and $T_ExcludePoor$, subjects are simply informed about the vote aggregation process. Subjects in all treatments are informed that the 99 other participants see the same information, but are *not* informed about the number of votes being excluded or manipulated by these interventions. Following the referendum, each individual i decides whether she wants to $Give_i|Rule:Give \in \{0, 1\}$ conditional on $Rule:Give$ being elected and whether she wants to $Give_i|Rule:Don't \in \{0, 1\}$ conditional on $Rule:Don't$ being elected (strategy method). Round 2 ends with an incentivized elicitation of beliefs about the choices of the other 99 participants in their session. After all participants have finished the experiment, random draws are executed, subjects are matched into pairs and decisions are being implemented. Subjects receive all payments and an e-mail with a summary of the outcomes within two days after the experiment. Figure 3.4 summarizes the timeline of an experimental session.

Belief Elicitation. In all treatments, at the end of round 2, we ask participants to state their beliefs about how many of the other 99 group members (a) follow *Rule:Give* (b) follow *Rule:Don't* and (c) vote for *Rule:Give*. We incentivize truth telling by letting subjects indicate a bracket (0-9 subjects, 10-19 subjects...,..., 90-99 subjects) and paying them GBP 0.50 for each question where the true number of subjects falls into this bracket (see Schlag and Tremewan, 2016, for a discussion of this method). In *T_Pay4Vote*, *T_Bribe* and *T_ExcludePoor*, we also elicit beliefs about the impact of the intervention on final voting outcomes. In *T_ExcludePoor*, we ask participants to state their belief about the share of votes for *Rule:Give* separately for the high income (income > GBP 40,000) and for the low income participants (income \leq GBP 40,000). In *T_Pay4Vote* we ask participants to state their beliefs about the share of *Rule:Give*-voters who pay for their vote and, separately, about the share of *Rule:Don't*-voters who pay for their vote. We do the same regarding the beliefs about the share of participants who accept the bribe in *T_Bribe*. Truth telling is incentivized in the same way as before, with subjects now indicating a bracket between 0-9% and 90-99%.

Post-Experimental Questionnaire. We conduct a post-experimental questionnaire to complement the standard background information on subjects we can access via Prolific.ac. The questionnaire is presented as an unrelated survey (using a different visual design and researcher profile) and is sent to the same people about two weeks after they participated in the experiment. These measures are meant to minimize the risk of spillovers from decision in the experiment and especially from exposure to the different treatments to questionnaire answers. We ask participants about their standpoints on various political issues such as redistribution, corruption, democratic values, and personal trust in institutions. Most of the questions are either directly taken or adapted from questions featuring in the 6th wave of the World Value Survey (WVS, 2014). Additionally, we assess personality characteristics such as risk preferences (self-reported and hypothetical lottery choice), trust, and the Big Five personality traits. The questions and answer format (7 point Likert scale) of the very short version of the Big Five are taken from Gosling, Rentfrow and Swann (2003). The full list of questions can be found in the appendix.

3.2.3 Empirical Strategy

To identify the impact of our interventions ($T_Pay4Vote$, T_Bribe , or $T_ExcludePoor$) on compliance, we cannot rely on comparing compliance rates in these treatments with the compliance rate in $T_Baseline$. Even though treatments are randomly assigned, treatment groups might differ in the ex-ante motivation of the average individual to follow a given rule. This can affect compliance levels and potentially hide or exaggerate treatment effects: Individual i may be more likely to follow a rule in the case that the rule corresponds to her individually preferred behavior or in the case that it corresponds to what she believes is the correct “societal” or “ought” behavior. We identify and control for these two motives by controlling for the type of an individual as indicated by her round 1 choice $Give_i \in \{0, 1\}$ and her $Vote_i \in \{0, 1\}$, indicating her preferred societal rule. Because treatment interventions are introduced after the votes are submitted in round 2, both variables are unbiased by the interventions. This identification is very close to the approach suggested by Dal Bó, Foster and Putterman (2010). Similar to them, we can estimate treatment effects on the type-level by conditioning on $Give_i \in \{0, 1\}$, $Vote_i \in \{0, 1\}$, or both. We go one step further and use the distribution of types in our experimental sample to estimate *average* treatment effects on the population level. Because there is no punishment associated with violating a rule, residual treatment differences measure to what extent the willingness to follow rules depends on the election process.

3.2.4 Implementation

The experiment is implemented online using a subject pool of (non-representative) international participants on the platform Prolific.ac based in Oxford, UK.¹⁷ We programmed the experiment using the software *LimeSurvey*, screenshots can be found in the appendix. All sessions were run in February and March 2017 on Tuesday, Wednesday or Thursday afternoons in order to keep the external circumstances as similar as possible between treatments. Registered participants have a unique Prolific-ID that is used to identify subjects, to prevent repeated participation and to process payments. When selecting into the experiment, *all* subjects see that they will take part in a lottery that pays GBP 100 to one out of 100 participants and that they will receive a fixed base payment of GBP 1.60 for completing the study.¹⁸ With each

¹⁷<https://prolific.ac>

¹⁸In the case of $T_Pay4Vote$ we increase the base payment by GBP 0.20 to counter adverse wealth effects when subjects pay to make their vote count. This is only announced after they

session taking roughly 15 minutes to complete, this base payment translates into an hourly wage of GBP 6.40. Additional payments are announced during the course of the experiment. For completing the 10 minute post-experimental questionnaire, subjects receive a compensation of GBP 1. The follow-up-rate is close to 100 per cent.¹⁹ In addition, subjects’ unique Prolific-ID allows us to access an extensive set of self-reported socio-demographic data including gender, nationality and income (see table 3.1). All information is provided voluntarily by the subjects but we required that only those who had filled out information on their gender and nationality were eligible for our study. For treatment *T_ExcludePoor* we also required that participants had filled out information on their annual household income (to make our intervention possible). To have a balanced sample in this particular treatment, we invited 50 participants with a stated household income above GBP 40,000—whose vote is counted in the election—and 50 participants with a stated household weakly below GBP 40,000—whose vote is *not* counted.²⁰ Table 3.1 shows a summary of sample demographics. With a mean age of 31, almost two thirds of the participants not being students and about one third having a non-Western nationality, our population sample differs in several respects from the typical subject pool at Western university labs.

	<i>Age</i>	<i>Female</i>	<i>Western</i>	<i>Student</i>	<i>Unemployed</i>	<i>UGrad</i>	<i>Inc < 40K</i>
Mean	31	0.42	.68	.36	.17	.58	.61
Std.Dev.	10.7						
Observations	394	400	400	400	400	390	321

Table 3.1: Participant Demographics. *Western* = 1 if Nationality is Western Europe, Australia, Canada, New Zealand, US. *Student* = 1 if participant is student at the moment of taking part. *UGrad* = 1 if highest education is at least undergraduate degree (BA/BSc/other). *Inc < 40K* if self-reported yearly household income is below GBP 40,000.

selected into the study, the base payment announced on the prolific website is the same across all treatments.

¹⁹Of 400 subjects, 387 filled out the questionnaire.

²⁰Individuals registered on *Prolific.ac* can access a list of active studies for which they are eligible and can participate in. They are *not* informed about the criteria used to pre-select “eligible” participants. For example, in treatment *T_ExcludePoor*, they do *not* know that eligibility is based on stated household income.

3.3 Treatment Effects

To set the stage for the analysis of treatment effects, we begin by providing summary statistics of choices that precede the compliance decisions of subjects as well as of the impact of our interventions on the voting outcome. We also provide an overview of subjects' beliefs about the behavior of other participants in their group. The overall giving rate in round 1—that is, the share of subjects choosing $Give_i = 1$ —is 61% (245/400).²¹ Almost all of those who choose to give in round 1 also vote for *Rule:Give* in the beginning of round 2 (93%). Among those who do not give in round 1, a significant majority of 59% vote for *Rule:Don't*. Overall, 73% vote in favor of *Rule:Give*, making it the preferred rule in every session. As a result of the treatment interventions, a considerable share of votes are either not counted or reversed: 35% of participants in *T_Pay4Vote* refuse to pay a fee to make their vote count, 39% of participants in *T_Bribe* accept to reverse their vote for the payment, and, by design, 50% of voters are excluded due to a low household income in *T_ExcludePoor*, see also Figure 3.5. Intuitively, excluding a substantial fraction of voters can affect the voting outcome. We measure “outcome bias” as the absolute value of the difference between the share of votes for *Rule:Give* before and after the intervention. While a large share of participants lose their voice, this has a relatively small effect on the voting outcome, see the right panel of Figure 3.5. In *T_Pay4Vote* the bias is in favor of *Rule:Give* (+5 percentage points), while in *T_Bribe* and *T_ExcludePoor* the bias is in favor of *Rule:Don't* (+11 and +3 percentage points, respectively). Beliefs about the impact of the treatment intervention (elicited at the end of the experiment) show that the large majority of subjects expected the interventions to lead to a considerable bias in the voting outcome (right panel of Figure 3.5).

Figure 3.6 shows the distributions of subjects' beliefs about the voting behavior and rule compliance of other participants in their session. From the histograms in the top panels we can see that beliefs are very heterogeneous. The median answer bracket regarding the question of how many of the other 99 participants voted for *Rule:Give* (panel a) is 50-59. This and the observation that the number of subjects stating extreme beliefs (0-9 or 90-99) is small gives us confidence that most subjects

²¹Note that our dictator game version differs in many respects from standard implementations of the game, namely by having ex-ante choices with role uncertainty, binary decisions, risky prospects with a small probability to win a high price, and by having an online participant sample. Still, the observation that 61% of subjects chose to give tickets does not deviate much from previous findings in the literature. For instance, in a meta-study of 129 dictator game studies covering 41,433 observations, Engel (2011, p.6) finds a share of 63.89% of subjects giving non-zero amounts.

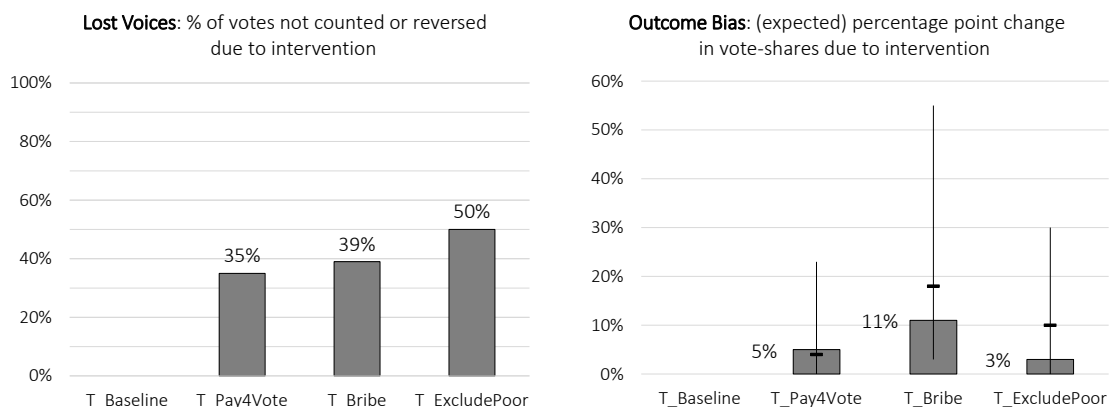


Figure 3.5: Left panel: Share of votes not counted or reversed in each treatment. Right panel: Outcome bias (absolute difference in the share of votes for *Rule:Give* before and after the intervention) in percentage points. The figure shows the actual outcome bias (as bars) as well as the distribution of subjects' beliefs about the outcome bias (median and 10th-90th percentile).

How many of the other 99 participants do you think...

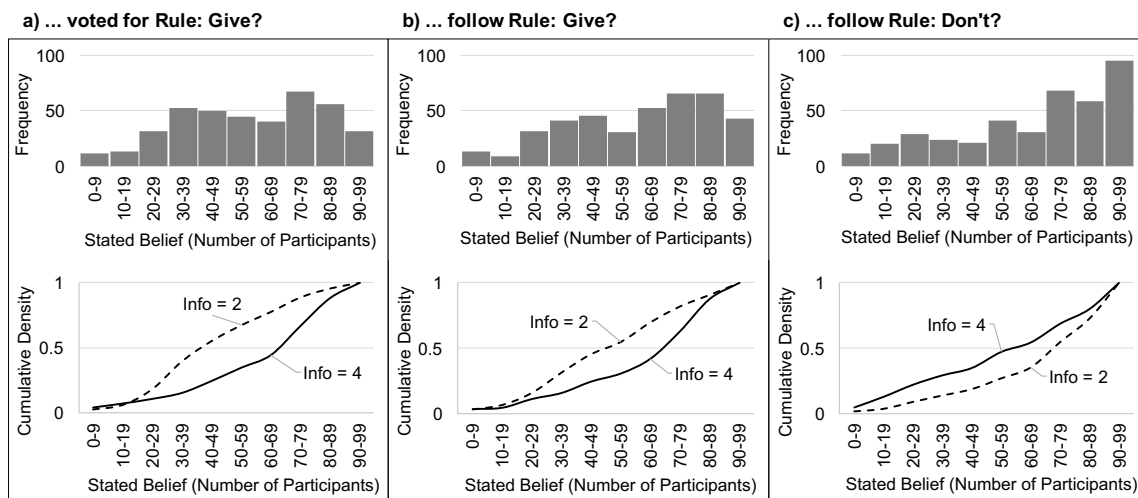


Figure 3.6: Beliefs about the choices of other participants (data from all treatments pooled, $N=400$). Top: Frequency of beliefs by answer bracket. Bottom: Cumulative density of answers among subjects having received *info=2* and *info=4*, respectively.

believed each of the two rules to have positive probability of being selected in the referendum. On average, subjects expect more people to comply with *Rule:Don't* (panel c) than with *Rule:Give* (panel b). The bottom graphs (cumulative densities) show that our information treatment was successful in shifting beliefs regarding the number of *Givers* in their group: among subjects who received the information that

four out of five subjects in an earlier study chose to give ($info=4$), beliefs about the number of participants voting for *Rule:Give* (panel a) and following *Rule:Give* (panel b) are consistently higher than among those subjects who received the information that only two out of five subjects chose to give ($info=2$). They also believe that less people choose to follow *Rule:Don't* (panel c).

3.3.1 Rule Compliance and Treatment Effects

Figure 3.7 delivers a first impression of the levels of rule-compliance with and without malpractice. The figure shows results separately for subjects who chose to *not* give in round 1 (*Non-Givers*, panel a) and those who chose to give in round 1 (*Givers*, panel b). Bar charts at the top of the figure depict compliance rates in the baseline treatment ($T_Baseline$). Here, we observe very high compliance rates: Almost every subject (98% of *Non-Givers* and 93% of *Givers*) follows the rule that prescribes the action that she preferred in round 1. More importantly, a significant fraction of subjects also follows the opposite rule: 65% of *Non-Givers* decide to follow rule *Rule:Give* and 53% of *Givers* decide to follow *Rule:Don't*. These numbers confirm a basic prediction of our model, namely that a democratically elected rule is voluntarily followed by more than just the original proponents of the action. As a consequence, overall giving rates in the baseline treatment react strongly to rules. The share of subjects who give increases from 57% in round 1 to 81% under *Rule:Give* and drops to only 28% under *Rule:Don't*.

Result 3.1 (Rule-Compliance without Malpractice). *In the absence of electoral malpractice, democratically elected rules have strong influence on voluntary behavior: Conditional on Rule:Give (Rule:Don't) being elected, 81% (72%) of subjects in T_Baseline voluntarily comply. 54% of subjects in T_Baseline are “rule-followers” who comply with either rule given its election.*

The bottom graphs in Figure 3.7 show percentage point differences between compliance rates in $T_Baseline$ and compliance rates in each of the treatments involving electoral manipulation. We immediately see strong and significant treatment effects among subjects whose individual choice in round 1 was to *not* give (*Non-Givers*, panel a): Of them, roughly 20-25 percent less can be convinced to follow *Rule:Give* if this rule is elected in the presence of a voting fee ($T_Pay4Vote$), monetary offers to vote differently (T_Bribe), or without the participation of low-income voters ($T_ExcludePoor$). The share of *Non-Givers* who can be identified as rule-compliers—

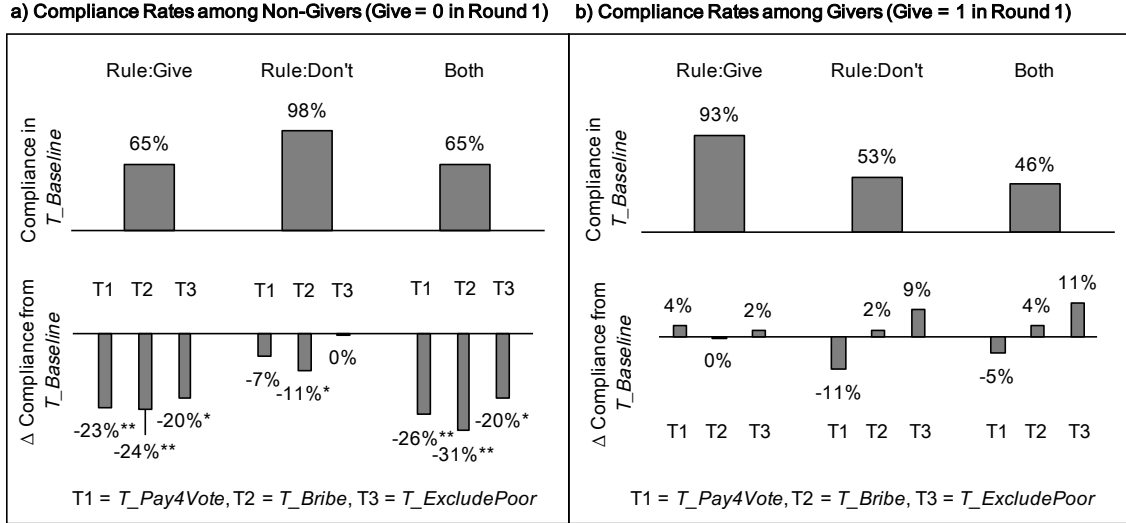


Figure 3.7: Compliance rates among a) *Non-Givers* (left panel) and b) *Givers* (right panel). $Both = 1$ if $Give_i | Rule:Give = 1$ and $Give_i | Rule:Don't = 0$. Top: Compliance rates in $T_{Baseline}$. Bottom: Percentage point change in compliance rates (Δ Compliance) from $T_{Baseline}$: T1 = $T_{Pay4Vote}$, T2 = T_{Bribe} , T3 = $T_{ExcludePoor}$. Stars denote significance level of one-sided Fisher-exact tests ($H_1: \Delta \text{ Compliance} > 0$): * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

those who voluntarily comply with either rule, if elected—drops from 65% without malpractice to only 34–45%. These responses are in line with our prediction that the manipulation of election processes lowers the utility to follow elected rules and thus diminishes voluntary rule-compliance. Maybe surprisingly, we find no evidence for such treatment effects being present among *Givers* (panel b): It seems that compliance with *Rule:Don't*—the rule we were expecting to see a deterioration in compliance among subjects who indicated a preference to give in round 1—is not affected by concerns about electoral manipulation.

To yield a deeper understanding of treatment differences and in order to calculate population average treatment effects, we classify subjects by

$$Type_i = Give_i (\text{Round } 1) \times Vote_i \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

and estimate effects of electoral malpractice for each type separately using OLS regressions. We present results from this approach in Table 3.2:²²

²²We discussed the necessity to control for $Give_i (\text{Round } 1) \in \{0, 1\}$ and $Vote_i \in \{0, 1\}$ in the identification section 3.2.3. In Table 3.2 we also control for possible effects of exogenous information $info_i \in \{2, 4\}$. Controlling for $info_i$ avoids sampling bias when running estimations on the smaller samples defined by types: Figure 3.6 shows that $info_i$ influences beliefs about the share of *Givers* in the population. Via this belief channel, the information treatment might influence compliance

		Number of subjects (all treatments)					Share of subjects complying with					
		By $Give_i$ (Round 1)					Rule: Give			Rule: Don't		
<i>All Treatm.</i>	By			all	<i>T_Baseline</i>	By	By $Give_i$ (Round 1)		By $Give_i$ (Round 1)		By $Give_i$ (Round 1)	
	$Vote_i$	0	1			0	1	all	0	1	all	
		0	92	17		109		0	.57	.50	.56	.96
	1	63	228	291		1	.80	1	.95	1	.51	.63
	all	155	245	400		all	.65	.93	.81	.98	.53	.72

(c) Treatment Effects on Compliance Rates (vs. Baseline):

		Rule: Give			Rule: Don't		
		By $Give_i$ (Round 1)			By $Give_i$ (Round 1)		
		0	1	all	0	1	all
<i>T_Pay4Vote</i>	By						
	$Vote_i$						
	0	-0.15 (.14)	.59 (.)	-.04 (.13)	-.05 (.07)	-.63 (.)	-.15* (.08)
1	-.35** (.16)	-.04 (.03)	-.11** (.04)	-.09 (.08)	-.07 (.10)	-.08 (.08)	
all	-.24** (.11)	.01 (.04)	-.09* (.05)	-.06 (.05)	-.11 (.09)	-.09 (.06)	
<i>T_Bribe</i>	By						
	$Vote_i$						
	0	.00 (.15)	-.02 (.)	.00 (.14)	-.09 (.08)	.28 (.)	-.03 (.08)
1	-.57*** (.18)	-.04 (.03)	-.16*** (.05)	-.16* (.09)	.04 (.09)	-.01 (.08)	
all	-.23* (.12)	-.04 (.04)	-.11** (.05)	-.12** (.06)	.05 (.09)	-.01 (.06)	
<i>T_ExcludePoor</i>	By						
	$Vote_i$						
	0	-.16 (.14)	.17 (.)	-.11 (.13)	-.01 (.07)	.27 (.)	.03 (.07)
1	-.33* (.18)	-.02 (.03)	-.09* (.05)	.00 (.09)	.10 (.10)	.08 (.08)	
all	-.23** (.11)	.00 (.04)	-.09* (.05)	-.01 (.05)	.12 (.09)	.07 (.06)	
<i>Pooled</i>	By						
	$Vote_i$						
	0	-.12 (.11)	.21 (.)	-.07 (.10)	-.04 (.06)	.02 (.)	-.03 (.06)
1	-.41*** (.14)	-.03 (.03)	-.11*** (.04)	-.08 (.07)	.02 (.08)	.00 (.07)	
all	-.23*** (.09)	-.02 (.03)	-.10*** (.04)	-.06 (.04)	.02 (.08)	-.01 (.05)	

Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.2: Number of subjects (a), baseline compliance rates (b) and treatment effects (c) by $Type_i = Give_i$ (Round 1) \times $Vote_i$ as well as average treatment effects for the entire population. White cells in (c) show coefficients and standard errors from OLS regressions of binary treatment variables on the compliance of types to *Rule:Give* ($Give_i|Rule:Give = 1$) and *Rule:Don't* ($Give_i|Rule:Don't = 0$), respectively, controlling for $info_i$. Grey cells show estimates of average treatment effects when types are weighted by population shares according to table (a).

decisions. Although this is not a cause of concern in large samples—given that $info_i$ is individual randomly drawn from a uniform distribution—, deviations from uniformity in smaller samples might

Panel a) reports the number of subjects of each type in the experimental population. Panel b) reports baseline compliance rates (the share of compliant subjects in *T_Baseline*) conditional on *Rule:Give* being elected (left-hand side) and conditional on *Rule:Don't* being elected (right-hand side). Panel c) reports treatment effects: It shows estimates of the change in compliance rates when going from *T_Baseline* to a treatment with electoral malpractice. Here, we first report separate treatment effects for each of the three malpractice treatments (*T_Pay4Vote*, *T_Bribe*, and *T_ExcludePoor*). In the lowermost section of panel c) we then report a “generalized” malpractice effect by pooling these data.

White cells in Table 3.2 panel c) show how malpractice affects the compliance of each type. For instance, the first four cells in the top-left corner of panel c) report the effects of implementing a voting fee on compliance with *Rule:Give* (*T_Pay4Vote*): Compliance drops by 15 percentage points among *Non-Givers* who voted for *Rule:Don't*, by 35 percentage points ($p < 0.05$) among *Non-Givers* who voted for *Rule:Give* and by 4 percentage points among *Givers* who voted for *Rule:Give*. Only among the $n = 3$ *Givers* in *T_Pay4Vote* who voted for *Rule:Don't* we measure a positive (and clearly, insignificant) effect.²³ To arrive at population average treatment effects, which are reported in the grey cells of the same panel, we weight types by their share in the experimental population. For example, we calculate the population average treatment effect of bribing voters (*T_Bribe*, *Rule:Give*) as $(92/400) \cdot (.00) + (63/400) \cdot (-.57) + (17/400) \cdot (-.02) + (228/400) \cdot (-.04) = -.11^{**}$. Standard errors for weighted averages are calculated using the Delta method.²⁴

Overall, Table 3.2 reinforces the impression from Figure 3.7: Electoral malpractice significantly affects compliance with rules promoting redistribution (*Rule:Give*), but seems to have little impact on compliance with rules opposing it (*Rule:Don't*). Treatment differences for *Rule:Don't* are small and (mostly) insignificant across all types. When pooling malpractice treatments (panel c, lowermost section), the population average treatment effect on compliance with *Rule:Don't* is estimated to be basically zero (-0.01 , $p = 0.87$). In contrast, apart from type $(Give_i, Vote_i) = (1, 0)$ —who only constitute 4% of the population—all types consistently show (weakly) lower compliance with *Rule:Give* if the vote aggregation process is manipulated in one way or the other. Compliance of subjects who did not give in round 1 but indicated

bias the estimates of treatment effects.

²³We do not report standard errors or significance levels for *Givers* who vote for *Rule:Don't* due to the tiny sample sizes. For the same reason we do not attempt to interpret their behavior.

²⁴For example, the standard error for the average treatment effect we just calculated can be determined from $\sqrt{(92/400)^2 \cdot (.15)^2 + (63/400)^2 \cdot (.18)^2 + (17/400)^2 \cdot (.37)^2 + (228/400)^2 \cdot (.03)^2} = .05$

a preference for *Rule:Give*—that is, compliance of type $(Give_i, Vote_i) = (0, 1)$ —is most volatile to whether the group selects this rule by democratic means: Among these participants, the share of subjects who follow *Rule:Give* drops by 35 percentage points in *T_Pay4Vote*, 57 percentage points in *T_Bribe* and 33 percentage points in *T_ExcludePoor*. Across all subjects who did not give in round 1, treatment effects closely match the effects displayed in Figure 3.7 (-24, -23, and -23 percentage points, respectively). Weighting these types in the total population we estimate that all three forms of electoral malpractice significantly reduce the overall share of individuals complying with *Rule:Give* by roughly 10 percentage points ($p < 0.1$, $p < 0.05$). Note that all three treatments show very similar effects on compliance rates, both on the type- and the aggregate level. Pooling the data (panel c, lowermost section), treatment effects for *Rule:Give* are significant at the 1 percent level.

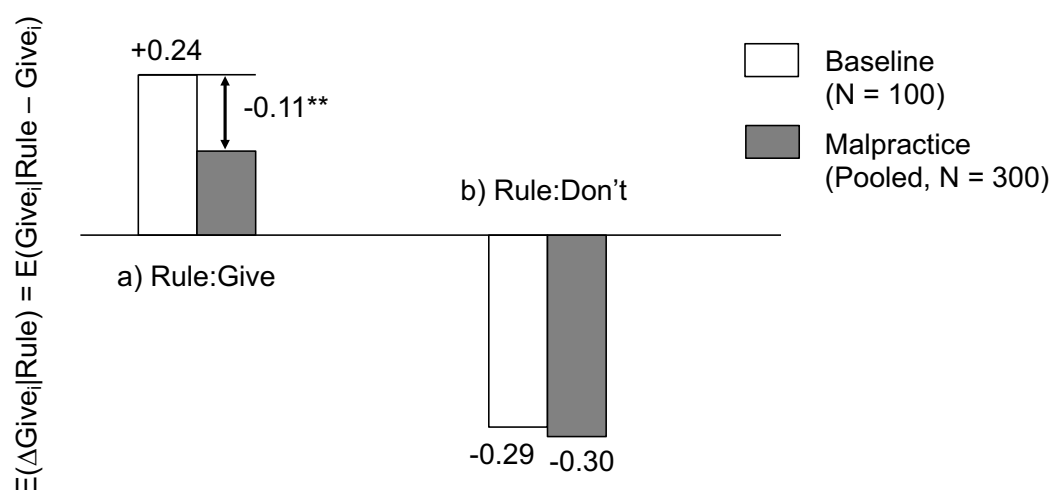


Figure 3.8: Power of the democratic vote to change individual behavior. Left-hand side (panel a): Average of $\Delta_i(Give|Rule:Give) := Give_i|Rule:Give - Give_i$. Right-hand side (panel b): Average of $\Delta_i(Give|Rule:Don't) := Give_i|Rule:Don't - Give_i$. Stars denote significance level of the coefficient on a binary treatment variable for malpractice (= 1 if individual i is in treatment *T_Pay4Vote*, *T_Bribe* or *T_ExcludePoor*) in a univariate OLS regression on $\Delta_i(Give|Rule:Give)$ (=Difference-in-Differences estimator). ** $p < 0.05$

Our analysis suggests that what is losing out under malpractice is the (non-coercive) power of a democratic vote to change individual behavior. A different way to look at the results is to make this loss in power explicit. Figure 3.8 shows the average difference between an individual's choice to give conditional on *Rule:Give* (*Rule:Don't*) being elected (round 2) and her choice before the referendum (round 1)—that is, the average of

$\Delta_i \text{Give} | \text{Rule:Give} := \text{Give}_i | \text{Rule:Give} - \text{Give}_i$ (Round 1) (on the left-hand side), and the average of $\Delta_i \text{Give} | \text{Rule:Don't} := \text{Give}_i | \text{Rule:Don't} - \text{Give}_i$ (Round 1) (on the right-hand side), respectively. If the democratic vote has power, one would expect *Rule:Give* to increase giving rates ($E(\Delta_i \text{Give} | \text{Rule:Give}) > 0$) and, conversely, *Rule:Don't* to decrease giving rates ($E(\Delta_i \text{Give} | \text{Rule:Give}) < 0$). This is also what we observe in the data. Consistent with our previous analysis, manipulations of the electoral process do not affect the power of *Rule:Don't*. *Rule:Give*, on the other hand, loses roughly half of its power to positively affect behavior. We summarize our findings regarding treatment effects below.

Result 3.2 (Main Result: Treatment Effects). *The manipulation of electoral processes significantly lowers voluntary compliance with Rule:Give. Of subjects who did not give before the election, on average 23 percent less ($p < 0.01$) can be convinced to follow Rule:Give in the presence of a voting fee (*T_Pay4Vote*), monetary offers to vote differently (*T_Bribe*), or without the participation of low-income voters (*T_ExcludePoor*). This translates into a 10 percentage points reduction of the compliance rate in the total population ($p < 0.01$) and is equivalent to the rule losing roughly half of its non-coercive power to change individual behavior. We find no evidence of electoral manipulation affecting compliance with Rule:Don't.*

3.4 Understanding Treatment Effects

What drives the strong adverse treatment effect on voluntary compliance with *Rule:Give*? Why is compliance with *Rule:Don't* not affected by manipulations of the electoral process? In this section, we will try to better understand the psychological determinants of rule compliance by analyzing the role of beliefs in driving behavior. In addition, we will exploit variance in the individual effects of the treatment interventions as well as information we obtained from the questionnaire about subject characteristics to account for individual heterogeneity and thus, better understand the behavioral pattern.

3.4.1 Beliefs about the Behavior of Other Subjects

We observe that rules have strong influence on voluntary behavior (see, for example, Figure 3.7). Do people follow rules because they want to follow others? Can this explain the treatment effects? Visually comparing the distribution of individual

beliefs about the behavior of other participants in treatment $T_Baseline$ with the respective distributions in treatments $T_Pay4Vote$, T_Bribe and $T_ExcludePoor$, we do not observe systematic differences.²⁵

Confirming this are the results of two-sample Kolmogorov-Smirnov tests which can also not reject equality of these distributions. This makes beliefs about others an unlikely candidate to explain treatment differences. Nonetheless, they may be an important determinant of rule-compliance in general: Understanding the causal effect of beliefs about others on the decision to comply with $Rule:Give$ and $Rule:Don't$, respectively, may help us explain the overall pattern of choices observed in the experiment.

Table 3.3 presents the results of an instrumental variable approach to estimating the role of others in guiding behavior under $Rule:Give$ (panel a) and $Rule:Don't$ (panel b). The main variable of interest in this analysis is $E_i(Comply_{-i}|Rule)$, which is the share of the 99 other participants whom individual i believes to be complying with $Rule:Give$ or $Rule:Don't$, respectively.²⁶ Because $E_i(Comply_{-i}|Rule)$ might be endogenous in a regression on $Give_i|Rule$, we instrument it with the binary variable $1.[info_i = 4]$. As Figure 3.6 shows, $info_i$ on average has a strong effect on $E_i(Comply_{-i}|Rule)$. Because it is exogenously randomized, it is a valid instrument.

Table 3.3 is structured as follows. Columns (1) present results of an OLS regression of $1.[info_i = 4]$, a dummy for malpractice,²⁷ and type controls $Give_i \times Vote_i$ on $E_i(Comply_{-i}|Rule:Give)$ (panel a) and $E_i(Comply_{-i}|Rule:Don't)$ (panel b), respectively. The small and insignificant coefficients on malpractice are in line with the Kolmogorov-Smirnov tests indicating that treatments did *not* systematically alter beliefs about the behavior of other subjects. At the same time, the large and highly significant coefficients on $1.[info_i = 4]$ confirm the observation from Figure 3.6: Going from $info_i = 2$ to $info_i = 4$ increases (decreases) an individual's belief about the share of participants complying with $Rule:Give$ ($Rule:Don't$) on average by 13 percentage points ($p < 0.01$). Variable $info_i$ is thus a powerful instrument to assess the causal effect of beliefs about the behavior of others on choices under both rules. Columns (2) report results of an OLS regression using the same explanatory

²⁵Figure 3.6 plots the distribution of these beliefs when pooling all four treatments. Beliefs in each individual treatment follow very much the same distribution.

²⁶We ask subjects to state their belief about the *number* of compliant others in their treatment. The response of individual i identifies a bracket, $E_i(\#Compliers_{-i}|Rule) \in \{0-9, 10-19, \dots, 90-99\}$. $E_i(Comply_{-i}|Rule)$ is the median of this bracket divided by 99. For example, if $E_i(\#Compliers_{-i}|Rule) = 40-49$, then the median is 44.5 and $E_i(Comply_{-i}|Rule) = 44.5/99 \approx 0.45$.

²⁷ $Malpractice = 1$ if individual i is a subject in treatment $T_Pay4Vote$, T_Bribe or $T_ExcludePoor$.

	(a) Rule: Give					(b) Rule: Don't				
	$E_i(\text{Comply}_{-i})$ (1) OLS	(2) OLS	(3) IV	(4) OLS	(5) OLS	$E_i(\text{Comply}_{-i})$ (1) OLS	(2) OLS	(3) IV	(4) OLS	(5) OLS
$\text{info}_i = 4$.13*** (.02)			-.04 (.04)	-.04 (.04)	-.13*** (.03)			-.11*** (.04)	-.09** (.04)
$E_i(\text{Comply}_{-i})$.46*** (.07)	-.32 (.30)				.51*** (.07)	.87*** (.33)		
$\text{Malpractice} = 1$	-.02 (.03)	-.10*** (.04)	-.11** (.05)	-.11*** (.04)	-.10** (.04)	.01 (.03)	-.02 (.05)	-.02 (.05)	-.01 (.05)	-.05 (.05)
Female_i										.00 (.05)
Risk_Seeking_i										.00 (.01)
$\text{Betrayal}_{-Aversion_i}$.04*** (.01)
Constant	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for Type_i	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Add. Controls										Yes
Observations	400	400	400	400	375	400	400	400	400	375

Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.3: The role of others in guiding behavior. $E_i(\text{Comply}_{-i})$ is individual i 's belief about the share of other participants complying with the rule. $\text{Malpractice} = 1$ if individual i is in treatment $T_Pay4\text{Vote}$, T_Bribe or $T_Exclude\text{Poor}$. IV regressions are 2SLS with $E_i(\text{Comply}_{-i})$ being instrumented by $1 \cdot [\text{info}_i = 4]$.

variables on compliance with *Rule:Give* (panel a) and *Rule:Don't* (panel b), respectively. The strong and highly significant coefficients on $E_i(\text{Comply}_{-i})$ show that beliefs about the behavior of others and individual compliance decisions are highly correlated. However, due to possible endogeneity, this correlation does not imply causality. For this reason, in columns (3), we use an IV (2SLS) estimator. Using $1.[\text{info}_i = 4]$ as an instrument for $E_i(\text{Comply}_{-i}|\text{Rule})$, we find strong evidence that beliefs about the behavior of others causally explain compliance with *Rule:Don't* (panel b). Specifically, a 1 percentage point increase in the expected share of others who comply is estimated to increase the probability of individual i to also comply and not give by 0.87 percentage points ($p < 0.01$). Accounting for this effect, no other explanatory variable is significant at the 5 percent level. Maybe surprisingly, we find no evidence that compliance with *Rule:Give* (panel a) is driven by similar motivations: $E_i(\text{Comply}_{-i})$ is insignificant for compliance with *Rule:Give* at any reasonable confidence level. Most importantly, irrespective of whether we control for beliefs about the behavior of others directly (column 2) or via instrument info_i (column 3), malpractice is identified to have virtually the same effect on rule-compliance as before, that is, reducing compliance with *Rule:Give* by approximately 10 percentage points in the total population while having no significant effect on compliance with *Rule:Don't*. These results imply that the drop in voluntary compliance with *Rule:Give* which we observe in the presence of electoral manipulation ($T_Pay4\text{Vote}$, T_Bribe or $T_Exclude\text{Poor}$) is not mediated by mean-variance shifts of beliefs about the behavior of others. On this hand, our results speak against a signaling theory of legitimacy. Rather, manipulations of electoral processes seem to directly impact the intrinsic motivation of individuals to follow *Rule:Give*. The analysis of *Rule:Don't* shows, on the other hand, that concerns regarding the process of rule selection may not necessarily be the prime drivers of compliance with *any* type of rule. Here, in stark comparison to *Rule:Give*, a strategic motivation to follow the behavior others is the dominant explanation. Given that beliefs about the behavior of other subjects do not vary significantly between treatments, this observation goes some way in explaining why malpractice does not significantly affect the share of subjects following *Rule:Don't*.

Columns (4) and (5) of Table 3.3 underline the robustness of our findings by presenting variations on the same scheme. Columns (4) present results of an OLS regression using info_i directly as an explanatory variable instead of using it as an instrument for $E_i(\text{Comply}_{-i})$. This way, we control for *any* systematic dependency

between individual behavior and beliefs about the share of pro-social agents in the population—which are shifted by $info_i \in \{2, 4\}$ —instead of specifically controlling for strategic complementarity in compliance. Columns (5) extend this analysis by including an extensive battery of individual characteristics and questionnaire answers as controls.²⁸ In both cases, our findings—in particular, regarding the effects of electoral manipulation (reflected in the coefficient on *Malpractice*) and the role of others in guiding behavior (now reflected in the coefficient on $info_i$)—are unchanged. We summarize our results below.

Result 3.3 (Beliefs about the Behavior of Other Subjects). *Beliefs about the behavior of other subjects causally explain voluntary compliance with Rule:Don't: A 1 percentage point increase in $E_i(Comply_{-i})$ increases the probability of the average subject to also comply with Rule:Don't by 0.87 percentage points ($p < 0.01$). We find no evidence of beliefs about others causally affecting voluntary compliance with Rule:Give. In particular, variance in the beliefs about other subjects cannot explain the observed adverse effects of electoral malpractice ($T_Pay4Vote$, T_Bribe , $T_ExcludePoor$) on compliance rates: Treatment effects are likely to be driven by a loss in the intrinsic motivation of individuals to follow the rule.*

3.4.2 Individual Disenfranchisement and Beliefs about the Outcome Bias

While treatments $T_Pay4Vote$, T_Bribe and $T_ExcludePoor$ differ in the particular form of electoral malpractice, they have in common that due to the intervention (a) many individuals lose their voice in the decision making process and (b) many individuals believe that the outcome of the referendum is biased compared to a fair majority vote (see Figure 3.5). Could it be that these two effects—being *personally* disenfranchised in the election and having doubts about the referendum's *overall* representativeness—are driving the loss in intrinsic motivation to follow *Rule:Give*?

²⁸ $Risk_Seeking_i$ is questionnaire-answer on 11-point Likert-scale to “Are you a person who is generally willing to take risks (10) or do you try to avoid taking risks (0)?”. $Betrayal_Aversion_i$ is questionnaire-answer on 11-point Likert-scale to “Do you think that most people would try to take advantage of you if they got the chance (10), or would they try to be fair (0)?”. Control for $Type_i$ includes $Give_i$ (Round 1), $Vote_i$, and $Give_i$ (Round 1) \times $Vote_i$. Additional controls in (5) are: $Western_i$, $Student_i$, $UGrad_i$, number of mistakes in control questions, factor variables measuring political and social values in questionnaire, as well as *Big Five* personality test measures on 7-point Likert scales. All controls not shown in the table are estimated to have small, insignificant effects ($p > 0.1$).

Let

$$Lost_Voice_i = \begin{cases} 1 & \text{if } i \text{ is in } T_Pay4Vote \text{ and } Accept_Pay_i = 0 \\ 1 & \text{if } i \text{ is in } T_Bribe \text{ and } Accept_Bribe_i = 1 \\ 1 & \text{if } i \text{ is in } T_ExcludePoor \text{ and } Income_i < 40K \\ 0 & \text{otherwise.} \end{cases}$$

Also, let $E_i[Outcome_Bias]$ be the belief of individual i about the absolute size of the outcome bias.²⁹ As shown in Figure 3.5, there is substantial heterogeneity between subjects regarding these two variables *within* each treatment. In Table 3.4 we test whether this variance captures the variance in compliance with *Rule:Give* that we observe between treatments.

The table presents results from OLS regressions of treatment dummies and controls on $Give_i|Rule:Give$, to which we successively add $Lost_Voice_i$ and $E_i[Outcome_Bias]$ as additional explanatory variables. Column (1) repeats our main finding that all three forms of malpractice ($T_Pay4Vote$, T_Bribe , and $T_ExcludePoor$) significantly reduce compliance with *Rule:Give*. Column (2) adds $Lost_Voice_i$ as an explanatory variable, column (3) adds $E_i[Outcome_Bias]$ as an explanatory variable, and column (4) adds both. Table 3.4 suggests that, indeed, (a) the experience of having one's voice not being counted in the referendum and (b) doubts about the overall representativeness of the election may be the underlying cause for the loss in intrinsic motivation: Including either of the two in the regression leads to a strong reduction in the size and significance of treatment effects. Including both in the regression basically wipes out the treatment effects observed for T_Bribe

²⁹ $Outcome_Bias$ is defined as the absolute difference between the share of votes for *Rule:Give* when counting the original votes of all 100 subjects (before the intervention) and the share of votes for *Rule:Give* that are finally counted in the referendum (after the intervention). The belief about the size of this bias is calculated from elicited beliefs with the following formula:

$$E_i[Outcome_Bias] := \begin{cases} 0 & \text{if } i \text{ is in } T_Baseline \\ \left| \frac{E_i[Accept_Pay_j | Vote_j = 1] E_i[Vote_j]}{E_i[Accept_Pay_j]} \right| & \text{if } i \text{ is in } T_Pay4Vote \\ |E_i[Accept_Bribe_j | Vote_j = 1] E_i[Vote_j] \\ + E_i[Accept_Bribe_j | Vote_j = 0] (1 - E_i[Vote_j])| & \text{if } i \text{ is in } T_Bribe \\ |E_i[Vote_j | Income_j > 40K] - E_i[Vote_j]| & \text{if } i \text{ is in } T_ExcludePoor \end{cases}$$

<i>Comply_i Rule:Give = 1</i>				
	(1)	(2)	(3)	(4)
	OLS	OLS	OLS	OLS
<i>Lost_Voice_i = 1</i>		-.11** (.04)		-.10** (.04)
<i>E_i[Outcome_Bias]</i>			-.34*** (.12)	-.33*** (.12)
<i>T_Pay4Vote</i>	-.11** (.05)	-.07 (.05)	-.08* (.05)	-.05 (.05)
<i>T_Bribe</i>	-.12** (.05)	-.08 (.05)	-.04 (.06)	.00 (.06)
<i>T_ExcludePoor</i>	-.09* (.05)	-.04 (.05)	-.06 (.05)	-.01 (.05)
Constant	Yes	Yes	Yes	Yes
Add. Controls	Yes	Yes	Yes	Yes
Observations	400	400	400	400
Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$				

Table 3.4: Explaining treatment variance in *Rule:Give* by variance in *Lost_Voice_i* (= 1 if individual *i*'s original vote is not counted in the referendum) and *E_i[Outcome_Bias]* ∈ [0, 1] (individual *i*'s subjective belief about absolute size of the outcome bias). OLS estimates. Regression includes constant and the following controls: *Give_i* (Round 1), *Vote_i*, *Give_i* (Round 1) × *Vote_i* and *info_i*.

and *T_ExcludePoor*. Only a small but insignificant effect remains for *T_Pay4Vote*.

Result 3.4 (Individual Disenfranchisement and Beliefs about the Outcome Bias). *Variance in Lost_Voice_i and E_i[Outcome_Bias] explains the variance between treatments: The experience of personally being disenfranchised in the election and having doubts about the referendum's overall representativeness may be underlying the loss in intrinsic motivation to follow Rule:Give that is observed in treatments T_Pay4Vote, T_Bribe, and T_ExcludePoor.*

3.4.3 Experience and Valuation of Democracy

Table 3.5 shows treatment effects separately for (1) subjects of western and non-western nationality, (2) subjects who state a high importance of living in a democratic country and those who do not, (3) subjects who claim to always participate in elections and those who do not, and (4/5) subjects who indicate a low justifiability for bribes and lobbying activities in the political sphere and those who do not. Information on nationality is provided to us by the survey platform (prolific.ac). Data

for the separation in Columns (2) to (5) comes from our questionnaire.

Table 3.5 suggests that our treatments may have affected a psychological domain that is associated with judgements of real world institutions: Significant treatment effects are found only among individuals who are likely to live in established democracies (column 1), who value democratic institutions (columns 2-3) and who strongly condemn violations of democratic principles (columns 3-4). Column (4) provides maybe the strongest support for this claim: Those who indicate a very high sensitivity to bribery in the real world also react very sensitively to electoral malpractice in our experiment. Those who find the acceptance of bribes in the course of one's duties at least sometimes acceptable, on the other hand, show only small and insignificant responses.

Result 3.5 (Experience and valuation of democracy). *The adverse effect of malpractice on compliance with Rule:Give is strong and significant only (1) among subjects who have a Western nationality, (2) among subjects who self-identify to value democratic institutions highly and (3) among subjects who indicate a low justifiability for bribes and (political) lobbying in the real world.*

3.5 Conclusion

We have presented the results of an online experiment that allows us to causally estimate how the introduction of a voting fee, monetary incentives to change voting behavior or the exclusion of poor voters from the ballot affect compliance with elected rules of behavior in a dictator game. Our results show that such attempts at manipulating a democratic voting process can have strong and significant adverse effects on the willingness of people to follow rules promoting redistribution (*Rule:Give*). We conclude that electoral malpractices, which are prevalent in many countries around the world, may undermine the positive effects of democracy on behavior that earlier research in public economics has established (see, for example, Frey, 1997; Tyran and Feld, 2006; Ertan, Page and Putterman, 2009; Sutter, Haigner and Kocher, 2010; Dal Bó, Foster and Putterman, 2010). Additional to this main result, our experiment provides insights into the psychological patterns underlying treatment effects and compliance behavior. We show that in our experiment, the adverse effects of vote buying and partial disenfranchisement on compliance cannot be explained by variance in beliefs about other participants' behavior. Rather, subjects seem to react intrinsically to violations of inclusiveness and unbiasedness in democratic elections.

		<i>Comply_i</i> Rule: Give = 1									
		Demographics		Questionnaire Data							
		(1)		(2)		(3)		(4)		(5)	
		Western Nationality		Importance of Democracy		Always participates in Elections		Justifiability of Bribes		Justifiability of Lobbying	
		Yes	No	High	Low	Yes	No	Low	High	Low	High
<i>T_Pay4Vote</i>		-.14** (.06)	-.08 (.08)	-.13* (.07)	-.10 (.07)	-.14** (.07)	-.06 (.08)	-.16** (.07)	-.06 (.08)	-.17** (.08)	-.06 (.07)
<i>T_Bribe</i>		-.12* (.06)	-.12 (.09)	-.18** (.08)	-.06 (.07)	-.15** (.07)	-.10 (.07)	-.22*** (.07)	-.02 (.08)	-.18** (.08)	-.09 (.07)
<i>T_ExcludePoor</i>		-.11* (.06)	-.01 (.09)	-.08 (.08)	-.10 (.07)	-.15** (.07)	-.05 (.07)	-.13* (.07)	-.06 (.08)	-.12 (.08)	-.07 (.07)
Constant		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Add. Controls		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations		272	128	183	201	190	194	214	170	205	179

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.5: Treatment effect heterogeneity by nationality and questionnaire responses to the following questions: (2) “How important is it for you to live in a country that is governed democratically?” (High = 10, Low = 1-9); (3) “When elections take place, do you vote always, usually, or never?” (Yes = always, No = other); (4,5) “Please indicate to what extent you think the following actions can be justified: (4) Accepting a bribe in the course of one’s duties. (5) Lobbying politicians to influence legislation.” (Low = can never be justified, High = other). OLS estimates. Regression includes constant and the following controls: $Give_i$ (Round 1), $Vote_i$, $Give_i$ (Round 1) \times $Vote_i$ and $info_i$.

This connects to earlier literature in psychology and behavioral economics which suggests that procedural aspects of decision making affect preferences directly (Tyler, 1990; Frey, Benz and Stutzer, 2004; Cappelen et al., 2013; Bartling, Fehr and Herz, 2014, among others). Interestingly, we find no evidence for our treatments affecting the willingness of people to comply with rules *opposing* redistribution: Compliance with *Rule:Don't* is high both in the presence and absence of electoral malpractice. Moreover, in stark contrast to behavior under *Rule:Give*, beliefs about the behavior of others are in this case a very strong causal determinant of compliance. It seems that rules demanding subjects to behave egoistically—maybe because such rules are less prevalent in the real world and thus, subjects are less familiar with such demands—trigger psychological responses that make the wish to follow others weigh stronger than concerns regarding the procedure of rule selection. It remains to be shown by future research whether this observation is robust and generalizable.

We consider our results to be of interest to several neighboring fields of literature. The observation that a majority of subjects in our experiment voted for the rule that is in line with their previous action yields insights into the relationship of private giving decisions and preferences over related social rules as discussed, for example, by Corneo and Grüner (2000, 2002). By showing that democratically elected, non-binding rules can impact people's propensity to act in a pro-social way we add insight to how norms in giving behavior (e.g. Krupka and Weber, 2013), inequality acceptance (e.g. Almås et al., 2010) and defaults for donations (e.g. Altmann et al., 2014) may be shifted and mediated in society. A generalization of our main result would suggest that people are less likely to follow pro-social rules (for example, to be honest) when these rules are advocated by a corrupt authority (in our case a flawed election). This provides one possible explanation for the observation made in earlier experiments (see, for example, Gächter and Schulz, 2016) that the level of corruption in a society is correlated with measures of individual intrinsic honesty: Living in societies with high levels of corruption might undermine the trust in institutions per se and thus, lead people to behave dishonestly even in unrelated experimental situations. Whether electoral manipulation is indeed associated with such a ripple effect is an exciting question for future research. Finally, our finding that behavior under *Rule:Don't* is strongly driven by a wish to follow the behavior of others, while behavior under *Rule:Give* is largely immune to such “peer effects” resonates with previous research on the contagion of pro-social and anti-social behaviors by Offerman (2002), Croson and Shang (2008), Thöni and Gächter (2015) and Dimant

(2017). Because pro-social behaviors are difficult to induce by peer-pressure, these studies have drawn the conclusion that an individual's own moral code of behavior is the main driving force behind pro-social choices. Our results show that group interactions *can* increase pro-social behavior, albeit not by appealing to the behavior of others but by the democratic election of a pro-social code of conduct.

Of course, this essay can only be a first step towards understanding the effects of electoral malpractice on behavior under democratically elected institutions. More research is needed to draw definitive conclusions. We chose to study rule compliance in the domain of redistribution for its important role in economic research and policy. However, we see our study primarily as making a claim about compliance to behavioral rules in general. Extending the analysis to other domains such as cheating and tax evasion as well as to other forms of centralized and de-centralized manipulation (such as ballot box stuffing and subject-to-subject bribes) is an important task for future research.

References

- Akerlof, Robert.** 2016. "Anger and Enforcement." *Journal of Economic Behavior and Organization*, 126: 110–124.
- Akerlof, Robert.** 2017. "The Importance of Legitimacy." *The World Bank Economic Review*, 30: 157–165.
- Almås, Ingvild, Alexander W. Cappelen, and Bertil Tungodden.** 2017. "Cutthroat Capitalism Versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking than Scandinavians?" HCEO Working Paper 2017-003.
- Almås, Ingvild, Alexander W. Cappelen, Erik Ø. Sørensen, and Bertil Tungodden.** 2010. "Fairness and the Development of Inequality Acceptance." *Science*, 328(5982): 1176–1178.
- Altmann, Steffen, Falk Armin, Paul Heidhues, and Rajshri Jayaraman.** 2014. "Defaults and Donations: Evidence from a Field Experiment." CESifo Working Paper No. 5118.
- Andreoni, James.** 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy*, 97(6): 1447–1458.

- Andreoni, James.** 1990. “Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving.” *The Economic Journal*, 100(401): 464–477.
- Bardhan, Pranab.** 2000. “Irrigation and Cooperation: An Empirical Analysis of 48 Irrigation Communities in South India.” *Economic Development and Cultural Change*, 48(4): 847–865.
- Bartling, Björn, Ernst Fehr, and Holger Herz.** 2014. “The Intrinsic Value of Decision Rights.” *Econometrica*, 82(6): 2005–2039.
- Basu, Kaushik.** 2015. “The Republic of Beliefs: A New Approach to ‘Law and Economics.’” World Bank Policy Research Paper No. 7259.
- Bénabou, Roland, and Jean Tirole.** 2012. “Laws and Norms.” *IZA Discussion Paper No. 6290*.
- Berman, Eli, Michael Callen, Clark Gibson, and James D. Long.** 2014. “Election Fairness and Government Legitimacy in Afghanistan.” NBER Working Paper No. 19949.
- Bernheim, B. Douglas.** 1994. “A Theory of Conformity.” *Journal of Political Economy*, 102(5): 8.
- Black, Sandra E., and Lisa M. Lynch.** 2001. “How to Compete: The Impact of Workplace Practices and Information Technology on Productivity.” *Review of Economics and Statistics*, 83(3): 434–445.
- Bolton, Gary E., and Axel Ockenfels.** 2000. “ERC: A Theory of Equity, Reciprocity, and Competition.” *The American Economic Review*, 90(1): 166–193.
- Bonin, John P., Derek C. Jones, and Louis Putterman.** 1993. “Theoretical and Empirical Studies of Producer Cooperatives: Will Ever the Twain Meet?” *Journal of Economic Literature*, 31(3): 1290–1320.
- Brusco, Valeria, Marcelo Nazareno, and Susan Carol Stokes.** 2004. “Vote Buying in Argentina.” *Latin American Research Review*, 39(2): 66–88.
- Callen, Michael, and James D. Long.** 2015. “Institutional Corruption and Election Fraud: Evidence from a Field Experiment in Afghanistan.” *American Economic Review*, 105(1): 354–381.

- Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden.** 2007. “The Pluralism of Fairness Ideals: An Experimental Approach.” *The American Economic Review*, 97(3): 818–827.
- Cappelen, Alexander W., James Konow, Erik Ø. Sørensen, and Bertil Tungodden.** 2013. “Just Luck: An Experimental Study of Risk-Taking and Fairness.” *American Economic Review*, 103(4): 1398–1413.
- Corneo, Giacomo, and Hans Peter Grüner.** 2000. “Social Limits to Redistribution.” *The American Economic Review*, 90(5): 1491–1507.
- Corneo, Giacomo, and Hans Peter Grüner.** 2002. “Individual Preferences for Political Redistribution.” *Journal of Public Economics*, 83(1): 83–107.
- Croson, Rachel, and Jen Shang.** 2008. “The Impact of Downward Social Information on Contribution Decisions.” *Experimental Economics*, 11(3): 221–233.
- Dal Bó, Pedro.** 2014. “Experimental Evidence on the Workings of Democratic Institutions.” In *Institutions, Property Rights, and Economic Growth: The Legacy of Douglass North.*, ed. Sebastian Gallani and Itai Sened. New York:Cambridge University Press.
- Dal Bó, Pedro, Andrew Foster, and Louis Putterman.** 2010. “Institutions and Behavior: Experimental Evidence on the Effects of Democracy.” *American Economic Review*, 100: 2205–2229.
- De Alth, Shelley.** 2009. “ID at the Polls: Assessing the Impact of Recent State Voter ID Laws on Voter Turnout.” *Harvard Law and Policy Review*, 3(1): 185–202.
- Dickson, Eric S, Sanford C Gordon, and Gregory A Huber.** 2015. “Institutional Sources of Legitimate Authority: An Experimental Investigation.” *American Journal of Political Science*, 59(1): 109–127.
- Dimant, Eugen.** 2017. “On Peer Effects: Contagion of Pro- and Anti-Social Behavior in Charitable Giving and the Role of Social Identity.” mimeo.
- Engel, Christoph.** 2011. “Dictator Games: A Meta Study.” *Experimental Economics*, 14(4): 583–610.

- Ertan, Arhan, Talbot Page, and Louis Putterman.** 2009. "Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free Rider Problem." *European Economic Review*, 53(5): 495–511.
- Fehr, Ernst, and Klaus M Schmidt.** 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 817–868.
- Fehr, Ernst, and Simon Gächter.** 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3): 159–181.
- Fischbacher, Urs, and Franziska Föllmi-Heusi.** 2013. "Lies in Disguise—An Experimental Study on Cheating." *Journal of the European Economic Association*, 11: 525–547.
- Frey, Bruno S.** 1997. "A Constitution For Knaves Crowds Out Civic Virtues." *The Economic Journal*, 107(443): 1043–1053.
- Frey, Bruno S., Matthias Benz, and Alois Stutzer.** 2004. "Introducing Procedural Utility: Not Only What, But Also How Matters." *Journal of Institutional and Theoretical Economics*, 160: 377–401.
- Gächter, Simon, and Jonathan F. Schulz.** 2016. "Intrinsic Honesty and the Prevalence of Rule Violations Across Societies." *Nature*, 531: 496–499.
- Gonzalez-Ocantos, Ezequiel, Chad Kiewiet De Jonge, Carlos Meléndez, Javier Osorio, and David W Nickerson.** 2012. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *American Journal of Political Science*, 56(1): 202–217.
- Gosling, Samuel D, Peter J Rentfrow, and William B Swann.** 2003. "A Very Brief Measure of the Big-Five Personality Domains." *Journal of Research in Personality*, 37(6): 504–528.
- Krupka, Erin L., and Roberto A. Weber.** 2013. "Identifying Social Norms Using Simple Coordination games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11(3): 495–524.
- Manza, Jeff, and Christopher Uggen.** 2008. *Locked Out: Felon Disenfranchisement and American Democracy*. Oxford University Press.
- Norris, Pippa.** 2014. *Why Electoral Integrity Matters*. Cambridge University Press.

- Offerman, Theo.** 2002. “Hurting Hurts More Than Helping Helps.” *European Economic Review*, 46: 1423–1437.
- Saito, Kota.** 2013. “Social Preferences under Risk: Equality of Opportunity versus Equality of Outcome.” *American Economic Review*, 103(7): 3084–3101.
- Schlag, Karl H., and James Tremewan.** 2016. “Simple Belief Elicitation.” *mimeo*.
- Sutter, Matthias, Stefan Haigner, and Martin G Kocher.** 2010. “Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations.” *The Review of Economic Studies*, 77(4): 1540–1566.
- Thöni, Christian, and Simon Gächter.** 2015. “Peer Effects and Social Preferences in Voluntary Cooperation: A Theoretical and Experimental Analysis.” *Journal of Economic Psychology*, 48: 72–88.
- Tyler, Tom.** 1990. *Why People Obey Rules*. Yale University Press.
- Tyler, Tom R.** 2006. “Psychological Perspectives on Legitimacy and Legitimation.” *Annual Review of Psychology*, 57: 375–400.
- Tyran, Jean-Robert, and Lars P. Feld.** 2006. “Achieving Compliance when Legal Sanctions are Non-Deterrent.” *The Scandinavian Journal of Economics*, 108(1): 135–156.
- Weber, Max.** 1978. *Economy and Society*. Berkeley: University of California Press.
- WVS.** 2014. “WORLD VALUES SURVEY Wave 6 2010-2014 OFFICIAL AGGREGATE v.20150418.” World Values Survey Association (www.worldvaluessurvey.org). Aggregate File Producer: Asep/JDS, Madrid SPAIN.
- Zwick, Thomas.** 2004. “Employee participation and productivity.” *Labour Economics*, 11(6): 715–740.

Appendix to Chapter 3

Theoretical Predictions for Voting Behavior

We extend our theory in Section 3.2 to yield predictions about voting behavior. Note that in all treatments, subjects vote before interventions take place that may undermine the democratic election. Voting decisions are therefore unbiased by the exposure to a particular treatment. We assume that each subject votes *sincerely* in the sense that she chooses to vote for the outcome that yields her a higher expected utility. Let $U_i[Rule]$ denote i 's expected utility given $Rule \in \{Rule:Give, Rule:Don't\}$. When voting, individual i takes into account how her own giving behavior will be affected by the rule as well as how the behavior of *other* subjects will be affected. Conditional on i not receiving tickets from the computer (which happens with probability 0.5), let $\Delta u(Receive) > 0$ denote the difference in utility between receiving three tickets from another subject and not receiving any tickets. Because the average subject in the population is more likely to give under *Rule:Give* than under *Rule:Don't*, the conditional probability that i will receive three tickets from another subject increases by

$$\Delta F[\bar{u}^D] = F[+\bar{u}^D] - F[-\bar{u}^D]$$

when going from *Rule:Don't* to *Rule:Give*. In our setup, voting behavior depends on the individual's giving preferences $\Delta u_i(Give)$ as follows:

1. *Unconditional Givers:* If $\Delta u_i(Give) \geq +\bar{u}^B$, individual i will choose $Give_i|Rule = 1$ irrespective of the rule. Individual i will then vote for *Rule:Give* ($Vote_i = 1$) if and only if

$$U_i[Rule:Give | (Give_i|Rule = 1)] \geq U_i[Rule:Don't | (Give_i|Rule = 1)]$$

$$0.5 \cdot [u_i(1) + \bar{u}^B] + 0.5 \cdot \Delta F[\bar{u}^B] \cdot \Delta u_i(Receive) \geq 0.5 \cdot u_i(1)$$

$$\Leftrightarrow \bar{u}^B \geq -\Delta F(\bar{u}^B) \cdot \Delta u(Receive).$$

2. *Unconditional Non-Givers:* If $\Delta u_i(Give) < -\bar{u}^B$, individual i will choose $Give_i|Rule = 0$ irrespective of the rule. Individual i will then vote for *Rule:Give* ($Vote_i = 1$) if and only if

$$U_i[Rule:Give | (Give_i|Rule = 0)] \geq U_i[Rule:Don't | (Give_i|Rule = 0)]$$

$$0.5 \cdot u_i(0) + 0.5 \cdot \Delta F[\bar{u}^B] \cdot \Delta u_i(\text{Receive}) \geq 0.5 \cdot [u_i(0) + \bar{u}^B]$$

$$\Leftrightarrow -\bar{u}^B \geq -\Delta F(\bar{u}^B) \cdot \Delta u(\text{Receive}).$$

3. *Rule-Followers*: If $-\bar{u}^B \leq \Delta u_i(\text{Give}) < +\bar{u}^B$, individual i will choose $\text{Give}_i(\text{Rule}) = 1$ under *Rule:Give* and $\text{Give}_i(\text{Rule}) = 0$ under *Rule:Don't*. Individual i will then vote for *Rule:Give* ($\text{Vote}_i = 1$) if and only if

$$U_i[\text{Rule:Give} | (\text{Give}_i | \text{Rule} = 1)] \geq U_i[\text{Rule:Don't} | (\text{Give}_i | \text{Rule} = 0)]$$

$$0.5 \cdot [u_i(1) + \bar{u}^B] + 0.5 \cdot \Delta F[\bar{u}^B] \cdot \Delta u_i(\text{Receive}) \geq 0.5 \cdot [u_i(0) + \bar{u}^D]$$

$$\Leftrightarrow \Delta u_i(\text{Give}) \geq -\Delta F(\bar{u}^B) \cdot \Delta u(\text{Receive})$$

We can see that there is a monotonic relation between $\Delta u_i(\text{Give})$ and the tendency to vote for *Rule:Give*. *Givers* always vote for *Rule:Give*. This is true for both, unconditional givers and rule-followers. If $\Delta F[\bar{u}^B]$ is close to zero, *Non-Givers* also vote according to their “natural” preferences, that is, $\text{Vote}_i = 0$. This case is illustrated in Figure 3.9, panel a). Increasing $\Delta F[\bar{u}^B]$ shifts voting preferences of non-givers in favor of *Rule:Give*. This first affects “moderate” *Non-Givers* who indeed would choose to give under the pro-social rule, i.e., those individuals who satisfy $-\bar{u}^B \leq \Delta u_i(\text{Give}) < 0$), see Figure 3.9, panel b). Only once $\Delta F[\bar{u}^B] \geq -\Delta \bar{u}^B / (\Delta u(\text{Receive}))$, also unconditional non-givers (and thus, all individuals) vote for *Rule:Give*, see Figure 3.9, panel c).

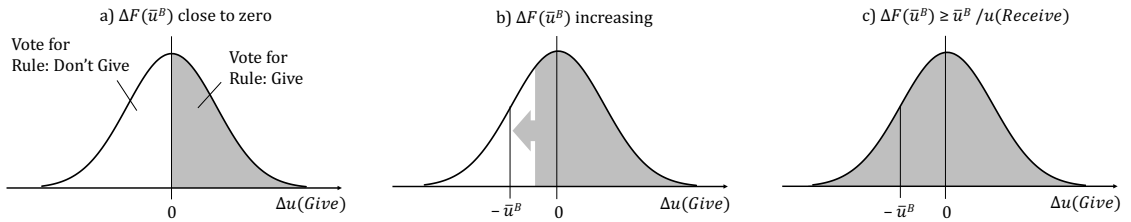


Figure 3.9: Theory: Share of Population voting for Rule: Give

Additional Data

	<i>T_Baseline</i>		<i>T_Pay4Vote</i>		<i>T_Bribe</i>		<i>T_ExclPoor</i>		<i>Pooled</i>	
Observations of which	100		100		100		100		400	
<i>Give_i</i>	= 0	= 1	= 0	= 1	= 0	= 1	= 0	= 1	= 0	= 1
Observations	43	57	43	57	29	71	40	60	155	245
<i>Info_i = 4</i>	.51	.42	.63	.39	.52	.49	.45	.53	.53	.46
<i>Vote_i = 1</i>	.35	.86	.47	.97	.45	.96	.38	.93	.41	.93
<i>Don't_Pay_i = 1</i>			.53	.21						
<i>Accept_Bribe_i = 1</i>					.76	.24				
<i>Excl_Poor_i = 1</i>							.48	.52		
<i>Give_i Rule:Give = 1</i>	.65	.93	.42	.97	.41	.93	.45	.95	.49	.94
<i>Give_i Rule:Don't = 0</i>	.98	.53	.91	.42	.86	.55	.98	.62	.94	.53
<i>Rule_Complier_i</i>	.65	.46	.40	.40	.35	.49	.45	.47	.47	.46

Table 3.6: Summary of experimental data. *Don't_Pay_i* = 1 if subject did not pay to make her vote count. *Accept_Bribe_i* = 1 if subject accepted to change her vote against payment. *Excl_Poor_i* = 1 if subject's vote was not counted because her stated household income is below 40.000 GBP. *Rule_Complier_i* = 1 if subject complies with both rules, i.e., *Give_i(Rule:Give)* = 1 and *Give_i(Rule:Don't)* = 0.

Questionnaire

Questionnaire: Politics

Overall, there are 15 questions. The first 10 questions relate to your views on politics.

1. In political matters, people talk of “the left” and “the right”. On a scale from 0 to 10, where would you place your views, generally speaking?

(Scale: 0 = Left, 10 = Right)

2. On a scale from 0 to 10, how important is it for you to live in a country that is governed democratically?

(Scale: 0 = not at all important, 10 = extremely important)

3. How democratic do you think your country is overall?

(Scale: 0 = not at all democratic, 10 = completely democratic)

4. How important is it for you to personally express your voice when it comes to political decision making?

(Scale: 0 = not at all important, 10 = extremely important)

5. It is important that you pay attention to this study. Please tick number 7 to show that you pay attention. The scale below does not play a role.

(Scale: 0 = not at all important, 10 = very important)

6. On a scale from 0 to 10, where 0 means “no trust at all” and 10 means “very much trust”, how much do you personally trust...

...politicians?

...large corporations?

...the results of elections?

7. Please indicate for each of the following actions to what extent you think that action can be justified:

(Scale: 0= can never be justified, 10= can always be justified)

- Violating the instructions of one’s superiors (for example at work or school).

- Accepting a bribe in the course of one's duties.
- Cheating on taxes if one has the chance.
- Influencing the actions of people by giving them money.
- Lobbying politicians to influence legislation.

8. Below you find two opposing statements on redistribution. How would you place your personal standpoint between the two statements (*0 means that you agree completely with the statement on the left, 10 means that you agree completely with the statement on the right*)

0:	10:
"The rich have an obligation to subsidize the poor. If necessary, they have to be forced to do so."	"Everybody is responsible for himself. Forcefully taking from the rich to subsidize the poor is theft."

9. Below you find two opposing statements on inequality. How would you place your personal standpoint between the two statements (*0 means that you agree completely with the statement on the left, 10 means that you agree completely with the statement on the right*)

0:	10:
"For a society to be fair, the incomes of all people should be equal."	"There is nothing unfair in having more money than somebody else, no matter how large the difference."

10. When elections take place, do you vote always, usually, or never?

Never Rarely Usually Almost always Always

Questionnaire: General questions

These are the final 5 questions of our study. They concern your views in general and your personality.

1. How do you see yourself: Are you a person who is generally willing to take risks, or do you try to avoid taking risks?

(Scale: 0 = Completely unwilling to take risks, 10 = Very willing to take risks)

2. How much do you agree with the following statement: “Money brings out the worst in people.”?

(Scale: 0 = Do not agree at all, 10 = Agree completely)

3. Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?

(Scale: 0 = All people would try to be fair, 10 = All people would try to take advantage of you)

4. Assume that you had the opportunity to take part in the following gamble: There are 100 balls in an urn. Of these balls, 99 are black and 1 is red. One ball is randomly drawn from the urn. If it is red you win 1000 GBP. If it is black you win 0 GBP. What would be the maximal amount of money you would be willing to pay in order to take part?

Would be willing to pay at most... (dropdown menu with answer choices from 0 GBP to 20 GBP in steps of 1)

5. Here are a number of personality traits that may or may not apply to you. Please indicate to what extent you agree or disagree that these personality traits apply to you.

Note: You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

I see myself as...


- Extraverted, enthusiastic (NOT reserved or shy)
- Agreeable, kind (NOT quarrelsome or critical)
- Dependable, self-disciplined (NOT careless or disorganized)
- Emotionally stable, calm (NOT anxious or easily upset/stressed)
- Open to new experiences, creative (NOT conventional)

(Scale: 1 = Disagree strongly, 2 = Disagree moderately, 3 = Disagree a little, 4 = Neither agree nor disagree, 5 = agree a little, 6 = agree moderately, 7 = agree strongly)

Instructions and Screenshots

Welcome

This study is hosted by:

 Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG [<https://www.uni-hamburg.de/en.html>]

Thank you for participating in our study! Your participation is very important to our research. The study takes about 15 minutes to complete and we ask you to please finish the study in one sitting.

Please read the following consent form before continuing:

I consent to participate in this research study. I am free to withdraw at any time without giving a reason (knowing that any payments only become effective if I complete the study).

I understand that all data will be kept confidential by the researchers. All choices are made in private and anonymously. Individual names and other personally identifiable information are not available to the researchers and will not be asked at any time. No personally identifiable information will be stored with or linked to data from the study.

I consent to the publication of study results as long as the information is anonymous so that no identification of participants can be made.

The study has received approval from the Dean's Office of the University of Hamburg, Germany.

If you have any questions about this research, please feel free to contact us at experiments@wiso.uni-hamburg.de.

To proceed, please give your consent by ticking the box below:

I have read and understand the explanations and I voluntarily consent to participate in this study.

Figure 3.10: Screenshot: Welcome and Consent Form

General Instructions

Please read the following instructions *very* carefully before proceeding with the study.

- This study has 100 participants. You are one of them.
- Each participant receives a base payment of £1.50 for completing the study. During the study, you may choose to invest £0.20 of this money. The minimum payment any participant receives is £1.30 (as announced on prolific.ac).
- One participant will receive an extra cash prize of £100. The winner of this cash prize is determined by a lottery. The chance of a participant to win the lottery depends on how many lottery tickets he/she holds at the end of the study.
- The number of lottery tickets you receive depends partly on luck and partly on yours and other participants' choices during this study. The final number of lottery tickets a participant holds ranges from 0 to 10. Each lottery ticket has the same chance to be the winning ticket.
- The winner of the £100 cash prize will be drawn once all 100 participants have completed the study and will be notified one week from now at the latest. You receive all payments through your [Prolific.ac](https://prolific.ac) account.
- Completion of the study at normal pace should not take more than 15 minutes.

Please tick this box when you are done reading the information and want to proceed.

I have read the information and want to proceed.

Figure 3.11: Screenshot: General Instructions

The Lottery

There are two rounds in this lottery:

- In each round, 500 lottery tickets will be distributed among the 100 participants. One of these lottery tickets is the winning ticket. The winning ticket yields the holder of the ticket a cash prize of £100. The final distribution of lottery tickets depends partly on luck and partly on the choices you and other participants make.
- Once all participants have completed the study, one of the two rounds will be randomly drawn to determine the final distribution of lottery tickets among participants.
- This means: Only the ticket distribution of one of the two rounds will be used to determine each person's chances to win. Each round has the same chance to be selected (50%) and the selected round will be the same for all 100 participants. We will inform you about the result of the random draw after you have completed the study.
- You will begin with round 1 of the lottery on the next screen.

Please tick this box when you have read the instructions and want to proceed:

I have read the instructions carefully and want to proceed.

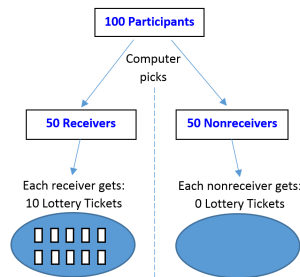
Figure 3.12: Screenshot: Instructions about the Lottery

Distribution of lottery tickets

In both rounds 1 and 2, the lottery tickets are distributed in two steps.

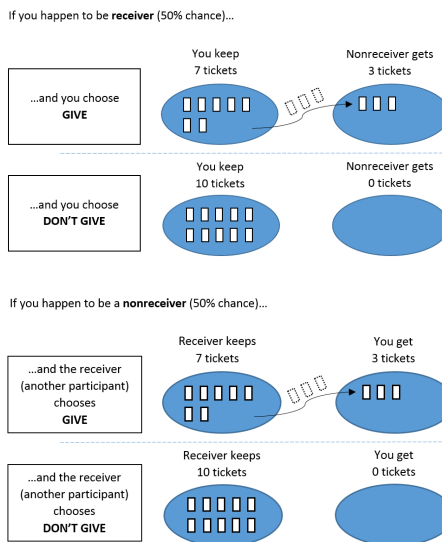
Step 1: The computer picks 50 receivers and 50 nonreceivers:

- The computer randomly selects 50 out of 100 participants to be "Receivers". Each receiver gets 10 lottery tickets from the computer.
- The other 50 participants are "Nonreceivers". Nonreceivers get 0 tickets from the computer.
- No participant learns whether he/she has been chosen to be a receiver or a nonreceiver until the end of the study.



Step 2: Participants decide whether they want to share tickets with nonreceivers:

- All participants decide—for the case they happen to be a receiver—whether they want to give 3 lottery tickets to a nonreceiver.
- This decision (GIVE or DON'T GIVE) has the following consequences:



When taking the decision whether to GIVE or DON'T GIVE, you will *not* know whether you have been selected to be a receiver or a nonreceiver. Nor will anybody else. You will receive a message with this information after all participants have finished the study.

If you happen to be a receiver (50% chance), your choice whether to GIVE or DON'T GIVE determines the final number of lottery tickets for you *and* for one other participant.

If you happen to be a nonreceiver (50% chance), your choice whether to GIVE or DON'T GIVE does *not* play a role. In this case, the choice of *another* participant (who happens to be a receiver) determines the number of lottery tickets that you will receive.

You will take the decision whether to GIVE or DON'T GIVE in both rounds 1 and 2.

Please make sure that you have understood the instructions given above. Once you are sure to have understood the instructions, please tick here to proceed.

I have read and understood the instructions and would like to proceed.

Figure 3.13: Screenshot: Instructions about the Distribution of Lottery Tickets

Round 1

Your Choice: Give or Don't Give

If you happen to be a receiver in round 1, do you want to GIVE or DON'T GIVE 3 of your 10 lottery tickets to a randomly selected participant who has received no tickets?

- We ask all participants to make this choice.
- If you happen to be a receiver, your choice will be automatically implemented.
- If you happen to be a nonreceiver, your choice does not play a role.
- Your choice remains private and anonymous to other participants.

Click here to be reminded of how lottery tickets are distributed to all participants of this study.

Remind me of the way lottery tickets are distributed.

Lottery tickets are distributed in two steps:

Step 1: The computer randomly selects 50 receivers and 50 nonreceivers. Each receiver gets 10 lottery tickets. Nonreceivers get no lottery tickets. No participant will learn whether he/she has been selected to be a receiver or a nonreceiver until the end of the study.

Step 2: Each participant decides privately whether he/she wants to GIVE or DON'T GIVE 3 lottery tickets to a nonreceiver for the case that he/she happens to be a receiver.

Please choose now:

GIVE 3 lottery tickets to a nonreceiver.

DON'T GIVE 3 lottery tickets to a nonreceiver.

Once you have made your decision, please tick below:

This is my final answer. Please proceed.

Figure 3.14: Screenshot: Choice $Give_i \in \{0, 1\}$ (Round 1)

End of Round 1

- Your choice in round 1 has been saved.
- You will be informed about the outcome of this round (whether you have been chosen to be a receiver or nonreceiver and how many lottery tickets you hold) via a private prolific.ac-message within one week of the end of this study.

Information about the choices of other people:

- To give you some information on how other people choose in the same situation, below you can see the choices of 5 participants *from an earlier study*:

Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
Don't Give	Give	Give	Don't Give	Don't Give

- Of these participants, 2 (out of 5) chose GIVE and 3 (out of 5) chose DON'T GIVE.

Please tick this box when you are done reading the information and want to proceed to round 2:

I have read the information and want to proceed to round 2.

Figure 3.15: Screenshot: Information $info_i \in \{2, 4\}$ (following Round 1)

Round 2

A code of conduct

In this round, lottery tickets will be distributed in the same way as in round 1.

Click here to be reminded of how lottery tickets are distributed to all participants of this study.

Remind me of the way lottery tickets are distributed.

Lottery tickets are distributed in two steps:

Step 1: The computer randomly selects 50 receivers and 50 nonreceivers. Each receiver gets 10 lottery tickets. Nonreceivers get no lottery tickets. No participant will learn whether he/she has been selected to be a receiver or a nonreceiver until the end of the study.

Step 2: Each participant decides privately whether he/she wants to GIVE or DONT GIVE 3 lottery tickets to a nonreceiver for the case that he/she happens to be a receiver.

However, before anyone decides anew whether to choose GIVE or DONT GIVE, a code of conduct will be set.

- The code of conduct says whether everyone should choose GIVE (\Rightarrow **RULE: GIVE**) or whether everyone should choose DONT GIVE (\Rightarrow **RULE: DONT GIVE**). Only one of the two rules will be implemented for this study.
- Once a rule has been set, all participants decide privately and anonymously whether they want to follow the rule or not.

Your vote: We ask each participant to vote for the rule (RULE: GIVE or RULE: DONT GIVE) he/she prefers to have implemented as the code of conduct for all participants. Please select a rule below.

Vote for RULE: GIVE

Vote for RULE: DONT GIVE

Once you have made your decision, please tick below:

This is my final answer. Please proceed.

Figure 3.16: Screenshot: $Vote_i \in \{Rule:Give, Rule:Don't\}$ (Round 2)

Round 2

Pay £0.20 to make your vote count

- You just selected RULE: DON'T GIVE as the rule you want to vote for.
- You have to pay £0.20 to make your vote count.

The code of conduct will be determined as follows:

- The rule that receives more votes in total will be implemented as the code of conduct.*
- The votes of participants who pay £0.20 will be counted. Other votes will not be counted.

*Tie Breaker: In case there are exactly the same number of votes counted for RULE: GIVE as for RULE: DON'T GIVE, a coin-flip decides which of the two rules will be implemented.

- If you pay £0.20, your vote for RULE: DON'T GIVE will be counted. If you don't pay, your vote will not be counted.
- This payment is independent of which rule you have selected (and whether or not the rule you have selected will be implemented).
- If you choose to pay, £0.20 will be subtracted from your base payment. All other payments are unaffected.
- We ask all 100 participants to make this choice. This means: Only the votes of those participants who pay £0.20 will be counted.

Please choose now:

Don't pay £0.20. Your vote will NOT be counted.

Pay £0.20. Your vote will be counted.

Once you have made your decision, please tick below:

This is my final answer. Please proceed.

Figure 3.17: Screenshot: $Accept_Pay_4Vote \in \{0, 1\}$ (Round 2, T_Pay_4Vote)

Round 2

Receive £0.20 for changing your vote

You just selected RULE: DON'T GIVE as the rule you want to vote for.

- The rule that receives more votes in total will be implemented as the code of conduct.*

*Tie Breaker: In case there are exactly the same number of votes counted for RULE: GIVE as for RULE: DON'T GIVE, a coin-flip decides which of the two rules will be implemented.

For an extra payment of £0.20: Are you willing to vote for the opposite rule instead?

- If you vote for the rule that is opposite to what you wanted to vote for (RULE: GIVE instead of RULE: DON'T GIVE), you will receive an extra payment of £0.20 on top of your base payment.
- This will be your final vote. Only the vote that you cast on this page will be counted.
- We ask all 100 participants to make the same choice. This means: All participants are offered an extra payment of £0.20 to vote for the rule that is *opposite to* what they originally wanted to vote for. Only the final vote of each participant will be counted.

Please choose now:

Accept extra payment of £0.20 and change my vote to RULE: GIVE.

Reject extra payment of £0.20 and keep my vote for RULE: DON'T GIVE.

Once you have made your decision, please tick below:

This is my final answer. Please proceed.

Figure 3.18: Screenshot: $Accept_Bribe \in \{0, 1\}$ (Round 2, T_Bribe)

Round 2

Your choice: Follow the rule or not

Your vote for the code of conduct has been counted.

▪ The rule that receives more votes in total will be implemented as the code of conduct.

Please choose now whether you want to follow the rule or not. Once a rule has been set, your choice for the relevant case will be automatically implemented.

If RULE: GIVE is implemented as the code of conduct, I choose to

Follow the rule and GIVE. Don't follow the rule and DON'T GIVE.

If RULE: DON'T GIVE is implemented as the code of conduct, I choose to

Follow the rule and DON'T GIVE. Don't follow the rule and GIVE.

Once you have made your decision, please tick below:

This is my final answer. Please proceed.

Figure 3.19: Screenshot: $Give_i | Rule \in \{0, 1\}$ (Round 2, $T_Baseline$)

Round 2

Your belief about other participants

Your choice has been saved and will be implemented accordingly.

As a final step, we are interested in your belief about the behavior of *other* participants in this round:

- All other participants make the same choices as you just did.
- For each question where your belief about the behavior of other participants is correct, you will receive an extra payment of £0.50 on top of your base payment. In total, you can earn up to £1.50 in extra payment on this page.

Click here to be reminded of how lottery tickets are distributed or of how the code of conduct is determined.

Remind me of how lottery tickets are distributed.

Remind me of how the code of conduct is determined.

How is the code of conduct determined?

- The rule that receives more votes in total will be implemented as the code of conduct.

1. How many of the other participants follow the rule?

a) If RULE: GIVE is implemented as the code of conduct, how many of the other 99 participants do you think follow the rule and GIVE?

	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

b) If RULE: DON'T GIVE is implemented as the code of conduct, how many of the other 99 participants do you think follow the rule and DON'T GIVE?

	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2. How do the other participants vote?

Of all other 99 participants, how many do you think have voted for RULE: GIVE to become the code of conduct?

	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Once you have made your decisions, please tick below:

These are my final answers. Please proceed.

Figure 3.20: Screenshot: Beliefs about Others (Round 2, *T_Baseline*)