IMT School for Advanced Studies Lucca
Lucca, Italy

Essays on Applied Behavioral Economic Theory

PhD Program in Systems Science
Track in Economics, Networks and Business Analytics
XXXVI Cycle

By

Bianca Sanesi

2026

The dissertation of Bianca Sanesi is approved.

PhD Program Coordinator: Massimo Riccaboni, IMT School for Advanced Studies Lucca

Advisor: Prof. Ennio Bilancini, IMT School for Advanced Studies Lucca

Co-Advisor: Prof. Federico Vaccari, University of Bergamo

The dissertation of Bianca Sanesi has been reviewed by:

Prof. Luca Corazzini, University of Milan-Bicocca

Prof. Andrea Gallice, University of Torino & Collegio Carlo Alberto

IMT School for Advanced Studies Lucca
2026

To my family

Contents

List of Figures

List of Tables

Acknowledgements

First of all, I would like to thank my advisor, Prof. Ennio Bilancini, for believing in me from the very beginning and for instilling in me the self-confidence to pursue research.

I am deeply grateful to my co-advisor, Prof. Federico Vaccari. His generosity of spirit and dedication have meant so much to me. Without him, I would not have been able to accomplish what I have. He is an example.

I would like to thank the reviewers, Prof. Luca Corazzini and Prof. Andrea Gallice, for their valuable comments and insights.

I am also grateful to all the professors and researchers who helped me during this PhD. Those who offered comments, suggestions, questions, or critiques. It is a long list of inspiring people. I will mention here my co-authors, Prof. Sibilla Di Guida and Prof. Leonardo Boncinelli, as well as my host during my inspiring visiting period at UPF, Prof. Larbi Alaoui.

I would like to thank all the members of the BEE group, especially Veronica.

A special thanks goes to my invaluable co-authors and friends, Nicco and Gine, who showed me the beauty of this work.

I would also like to thank my new family in Lucca. The friends I made while living at the IMT campus, whom I know I will never lose. Sharing this journey with you has taught me so much, not only as a researcher but as a person.

Thanks also to the friends in Firenze, among whom I know I can always count, and to Simo, who helped me in all ways possible.

I am deeply grateful to my mother, my father, and my sister, who have always supported my choices, even when they did not fully share them. I am grateful to my whole family, and in particular to my Nonna Paola, for her "preghierine," which seems to work even too well. As part of the family, I need to thank Elena. You're so special.

Finally, I want to thank Leli. Since you came into my life, you have brought light.

Vita

May 10, 1995	Born, Firenze (FI), Italy

Education

2020 – present	PhD, Economics, Networks, and Business Analytics IMT School for Advanced Studies Lucca Supervisor: *Ennio Bilancini* Co-supervisor: *Federico Vaccari*
2024 – 2025	Visiting PhD Student Pompeu Fabra University of Barcelona Host: *Larbi Alaoui*
2018 – 2020	M.Sc. Economics (LM-56) IMT School for Advanced Studies Lucca University of Florence Supervisor: *Leonardo Boncinelli* Grade: *110/110 cum Laude*
2016 – 2017	Erasmus programme IPAG Business School of Paris
2015 – 2018	B.A. Business Economics (L-18) University of Florence Supervisor: *Donato Romano* Grade: *110/110*

Work Experience

2025 – 202	Teaching Assistant *Microeconomics* by R. Ghidoni University of Bologna
2025 – 202	Teaching Assistant *Microeconomics* by F. Vaccari University of Bergamo
2023 – 2024	Teaching Assistant *Microeconomics* by S. Di Guida and K. Huremovic IMT School for Advanced Studies Lucca
2022 – 2023	Teaching Assistant *Microeconomics* by A. Canidio and K. Huremovic IMT School for Advanced Studies Lucca
2021 – 2022	Teaching Assistant *Information Economics* by F. Vaccari IMT School for Advanced Studies Lucca
2021 – 2022	Teaching Assistant *Game Theory* by E. Bilancini IMT School for Advanced Studies Lucca
2019 – 2020	Junior Tutor *Microeconomics* by N. Doni *Mathematics for Economics* by M. Gori University of Florence

Working Papers

1. Sanesi, Bilancini, and Boncinelli, 2024, "A model for rational generosity of the rich: Status concerns and poverty blaming when social classes are partially segregated,"
   Revise and Resubmit, *Journal of Economic Behavior and Organization*.
   Available at: `http://dx.doi.org/10.2139/ssrn.4946789`
2. Sanesi and Del Mastio, 2025, "The Shape of Moral Satisfaction,"
   Available at: `http://dx.doi.org/10.2139/ssrn.5503102`
3. Toccafondi, Sanesi, and Di Guida, 2025, "The demand effects of real-time er congestion disclosure,"
   Available at: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5591737`

Presentations

1. B. Sanesi, "Norms and Behavioral Change Talk (NoBeC)," at *Center for Social Norms and Behavioral Dynamics, University of Pennsylvania* , 19 March 2026.
2. B. Sanesi, "GREEN TIPPING Flyout Seminar (Job Market Seminar)," at *University of Bologna*, 6 February 2026.
3. B. Sanesi, "CLOSER Workshop on Experimental Social, Political and Economic Sciences," at *Collegio Carlo Alberto*, 26–27 January 2026.
4. B. Sanesi, "15th Meeting of the Behavioral and Experimental Economics Network (BEEN)," at *LUISS University*, Rome, 23 January 2026.
5. B. Sanesi, "Oikonomia – Relationships, Care, and the Economy," *Scuola universitaria professionale della Svizzera italiana (SUPSI)*, 2026, January 12.
6. B. Sanesi, "UCL Online Journal Club," at *University College London (UCL)*, Online, 31 October 2025.
7. B. Sanesi, "66ma Riunione Scientifica Annuale della Società Italiana di Economia," at *Università di Napoli Parthenope (SIE)*, 23–25 October 2025.
8. B. Sanesi, "1st SEEDS Summer School on Behavioural Economics for Sustainability," at *UnitelmaSapienza, Rome (SEEDS)*, 2–3 October 2025.
9. B. Sanesi, "CIMEO Workshop," at *Sapienza University of Rome – CIMEO Lab, Department of Economics and Law*, 16 September 2025.
10. B. Sanesi, "Subjective Probability, Utility, and Decision Making (SPUDM)," at *IMT School for Advanced Studies Lucca*, 31 August–4 September 2025.
11. B. Sanesi, "Summer School & Workshop on Experimetrics & Behavioral Economics," at *Sapienza Università di Roma (CIMEO)*, 19–24 July 2025.
12. B. Sanesi, "Annual Conference Society for the Advancement of Behavioral Economics (SABE)," at *University of Trento*, 5–7 June 2025.
13. B. Sanesi, "7th Meeting on Behavioural and Experimental Economics," at *University of Florence (BEELab)*, 8–10 May 2025.
14. B. Sanesi, "Microeconomic Reading Group," at *Universitat Pompeu Fabra, Barcelona (Microeconomics Department)*, 26 February 2025.
15. B. Sanesi, "UPF Student Seminar Series," at *Universitat Pompeu Fabra, Barcelona (PhD UPF)*, 17 October 2024.
16. B. Sanesi, "Summer School & Workshop on Experimetrics & Behavioral Economics," at *Sapienza Università di Roma (CIMEO)*, 21–27 July 2024.

17. B. Sanesi, "17th Workshop on Social Economy for Young Economists," at *University of Bologna (Forlì Campus)*, 13–14 June 2024.
18. B. Sanesi, "Annual PhD Meeting of the EADS and ENBA programs," at *IMT School for Advanced Studies Lucca (PhD Meeting)*, 4–6 February 2024.
19. B. Sanesi, "XVII GRASS Workshop," at *University of Florence (GRASS)*, 11–12 September 2023.
20. B. Sanesi, "22nd Society for the Advancement of Economic Theory Conference (SAET)," at *Paris Sorbonne*, 17–21 July 2023.
21. B. Sanesi, "12th IBEO - Alghero Political Economy," at *University of Sassari (IBEO)*, 3–4 July 2023.
22. B. Sanesi, "11th Oligo Workshop," at *University of Padua (OLIGO)*, 8–10 June 2023.
23. B. Sanesi, "Royal Economic Society & Scottish Economic Society 2023 Annual Conference (RES & SES)," at *University of Glasgow*, 3 April 2023.
24. B. Sanesi, "Meeting Young Economist of Tuscan Institutions (YETI)," at *IMT School for Advanced Studies*, 19 December 2022.
25. B. Sanesi, "European Meeting on Game Theory (SING17)," Online - *University of Padua*, 12 July 2022.
26. B. Sanesi, "Virtual Formal Theory Workshop (VFT)," Online - *American Political Science Association*, 4 April 2022.
27. B. Sanesi, "PhD Seminars," at *IMT School for Advanced Studies Lucca*, 14 January 2022.

Abstract

This dissertation advances applied behavioral economics theory from three complementary perspectives. First, it studies how concerns about social status, in the presence of social segregation, generate non-trivial incentives for redistribution. The analysis reveals that status competition can lead the wealthy to economically benefit the poor while simultaneously stigmatizing poverty. Second, it explores the welfare effects of competition in strategic communication, where lying is possible but costly. Relying on the behavioral assumption of the Principle of Insufficient Reason, the study formally demonstrates an intuitive result: competition benefits the receiver by improving information transmission. Third, it examines moral satisfaction, the hedonic reward from acting in line with one's moral values, as an additive component of utility. Through theoretical modeling and experimental evidence, the analysis identifies the shape of this moral satisfaction function and its role in explaining consistent and licensing moral behaviors. Taken together, these contributions highlight the importance of incorporating behavioral considerations into standard economic theory, revealing mechanisms and outcomes that remain hidden under traditional frameworks.

Introduction

Understanding human behavior often requires examining the intricate interplay between individual incentives, social environments, and the cognitive and emotional mechanisms that shape decision-making. This dissertation explores this interplay across three seemingly distinct domains: social status and redistribution, strategic communication, and moral behavior.

This research investigates how information becomes belief and how beliefs, both one's own and those of others, influence behavior. Concern for social status reflects the importance of others' beliefs about an individual's income. Strategic communication depends on how people update beliefs in response to messages, and moral concern arises from the value derived from acting in accordance with normative beliefs. By integrating behavioral insights into standard economic models, the analyses uncover richer and sometimes unexpected mechanisms. The first two chapters address the strategic challenges created by information asymmetries under behavioral assumptions, while the third extends the framework to moral behavior, demonstrating that traditional models of social preferences fail to capture the full depth of individuals' moral values.

In Chapter 1, the strategic dynamics of status competition and redistribution in a three-type, two-market signaling model are examined. Here, the interaction between different social types (poor, middle, and rich) generates unexpected incentives for the highest type. The partial segregation among types makes it advantageous for the rich to support

the poorest while undermining the middle, and to stigmatize poverty in ways that maximize their relative advantage. These theoretical insights offer explanations for real-world phenomena, including patterns of charitable giving, social segregation, and attitudes toward redistribution. The chapter highlights how heterogeneity in social types and market segregation shapes incentives and beliefs, providing a foundation for understanding social behavior.

Chapter 2 shifts to the domain of strategic communication, studying how competition among informed senders affects information transmission when lying is costly. While traditional theory predicts that competition always improves outcomes for receivers, costly misreporting changes the incentives of senders in subtle ways. To address this, we introduce a behavioral refinement, the Laplacian Criterion, grounded in the Principle of Insufficient Reason, which provides a tractable tool for equilibrium selection. Applying this refinement, we show that competition can indeed enhance receiver welfare even under costly communication. This chapter highlights how strategic interactions among senders and receivers shape the quality of information and decision-making outcomes, thereby extending our understanding of belief formation and, therefore, behavior.

In Chapter 3, the focus turns to moral behavior, specifically why individuals sometimes act consistently across moral decisions and at other times engage in moral licensing after a good deed. We argue that these patterns are driven by the curvature of moral satisfaction, the hedonic utility from acting in line with one's personal norms. The central question is whether individuals experience increasing or decreasing marginal moral satisfaction from good deeds, and how much this curvature varies across people. In our model, increasing marginal moral satisfaction leads to moral consistency, whereas diminishing marginal moral satisfaction results in moral licensing. We study this mechanism in an online two-stage experiment that holds material incentives constant while varying moral framing. Combining behavioral choices with linguistic measures of moral engagement allows for the recovery of substantial heterogeneity in the shape of moral satisfaction. We find that participants with

higher moral engagement exhibit diminishing marginal moral satisfaction, while less engaged participants display increasing marginal moral satisfaction. Finally, we find that moral satisfaction extends across domains: moral actions in different contexts appear to serve as substitutable sources of hedonic utility, helping explain when good deeds crowd in versus crowd out future prosocial behavior.

Taken together, the chapters demonstrate the central role of beliefs in mediating responses to incentives and social environments. Whether in social status, communication, or morality, accounting for heterogeneity and interaction structure enables a deeper understanding of behavior. The dissertation highlights the value of integrating theoretical modeling with behavioral reasoning to uncover the mechanisms underlying observed outcomes and to inform more effective policy design.

Chapter 1

Segregation, poverty stigma, and redistribution in a status signaling model

This chapter is based on Sanesi, Bilancini, and Boncinelli, 2024. Minor editorial assistance, limited to text refinement, was performed using AI-based tools.

1.1 Introduction

The desire to earn esteem within one's community shapes behavior across societies (Smith, 1759; Veblen, 1899). Because people care about how others perceive them (L. S. Bagwell and Bernheim, 1996; Bénabou and Tirole, 2006), they often pursue recognition through economically visible markers (Frank, 1985). This pursuit can induce conspicuous consumption: individuals spend on goods that communicate affordability, not only on goods that increase intrinsic enjoyment (Han, Nunes, and Drèze, 2010). Status concerns are inherently relative, since individuals compare themselves to a reference group and seek local distinction (Frank, 1985; Hopkins and Kornienko, 2004). Segregation matters in this context because it shapes who observes whom, and therefore which comparisons and signals become payoff-relevant.

Our analysis builds on an ordinal view of status. Individuals care about being recognized as belonging to a higher social class, not about the cardinal size of income or consumption gaps. This feature naturally turns status into an object of competition: because recognition is relative and categorical, individuals must choose signals that distinguish them from nearby types in their reference group. Conspicuous spending matters because it affects recognition, not because higher spending is intrinsically better.

We study a setting in which status competition is segmented across social classes. Poor and middle individuals interact in a low-status location, while middle and rich individuals interact in a high-status location. The middle class spans both arenas. We develop a three-type signaling model with this partial segregation and show that it can redirect status pressure away from the top. When competition between poor and middle intensifies, the middle class must spend more to separate downward, which leaves it with fewer resources to compete upward. This mechanism relaxes the rich–middle competition and can allow the rich to attain separation at lower cost. In this sense, the rich can gain by pushing status competition downward.

The mechanism relies on a specific feature of segregation: different strata use distinct and non-interchangeable symbolic languages of status. Signals that work in one location can lose meaning, or even backfire, in another. Segregation stabilizes these symbolic codes by restricting audiences and limiting cross-class visibility. This idea echoes Bourdieu (1979)'s observation that practices can become "out of place" outside the social space in which they are produced:

> *Jokes which 'fall flat' or, though acceptable in another context, in another market, here seem "out of place" and only provoke embarrassment or disapproval; quotations—in Latin, for example—which sound "pedantic" or "laboured."*

A simple illustration comes from clothing. Visible branding can function as an effective marker of affluence in lower-status environments, yet the same items may be perceived as vulgar in higher-status contexts

(Charles, Hurst, and Roussanov, 2009; Han, Nunes, and Drèze, 2010). Conversely, subtle signals (e.g., bespoke tailoring or minimalist design) can convey status among elites while remaining largely unrecognized in lower-status settings (Bellezza, Gino, and Keinan, 2014). Consistent with this view, individuals adjust consumption and behavior to match the symbolic codes used for status recognition in their reference group (O'Cass and McEwen, 2004).

We use the model to study the rich's incentives over redistribution. In the standard separating logic of signaling games, raising poor income can benefit the rich because it forces the middle class to spend more to distinguish itself from the poor and therefore weakens its ability to challenge the rich.[1] The same logic implies that narratives or institutions that stigmatize poverty can also benefit the rich: stigma raises the middle class's value of distancing itself from the poor and intensifies poor–middle competition. At the same time, the rich need not always face this outcome configuration. In some parameter regions, the rich must overspend to prevent the middle class from abandoning downward separation and reallocating resources to compete upward. Depending on the income-class definition, the middle share can be large but has declined in several OECD countries; in some settings, "below-middle + above-middle" is close to half of the population (Pew Research Center, 2015), making rich–poor coalitions empirically plausible. The model therefore delivers sharp conditions under which redistribution attracts elite support, and when it instead triggers stronger status expenditures at the top.

Relative to standard signaling models of conspicuous consumption, our framework introduces two linked departures. We consider three ordered social types rather than a single binary comparison, and we assume that status competition occurs in two partially segregated arenas rather than before a common audience. Poor and middle individuals interact in one arena, middle and rich individuals in another, and only the middle class spans both. This structure makes the middle class the channel through which stronger status pressure at the bottom affects compe-

[1] In a canonical separating outcome, the lower type spends zero on signaling while the higher type spends just enough to deter imitation (Riley, 2001).

tition at the top. As a result, the model delivers new predictions: redistribution toward the poor can benefit the rich by intensifying poor–middle separation; poverty stigma can advantage the rich by raising the middle class's incentive to distance itself from the poor; and partial segregation can reduce top-level conspicuous expenditure by shielding the rich from direct competitive pressure.

Finally, we show that the rich may prefer to promote partial segregation. By limiting cross-class visibility and keeping symbolic codes location-specific, segregation can lower aggregate conspicuous consumption through a decline in rich spending.

The chapter is structured as follows. Section 1.2 positions the contribution in the relevant literature. Section 1.3 presents the three-type, two-location signaling model. Section 1.4 characterizes equilibrium behavior. Section 1.5 explains how partial segregation reshapes standard predictions about status competition. Section 1.6 studies redistribution. Section 1.7 analyzes poverty stigma. Section 1.8 concludes.

1.2 Literature

Our paper relates to work on conspicuous consumption and status signaling, on the political economy of redistribution under status concerns, and on socioeconomic segregation and reference groups. A long tradition studies how individuals use consumption to display social standing, starting from Veblen (1899). Building on signaling theory (Spence, 1973a), a large literature models conspicuous expenditure as a costly signal of socioeconomic position and analyzes the resulting incentives to separate from nearby types in local status competitions (Frank, 1985; Norman J. Ireland, 1994; Hopkins and Kornienko, 2004). This line of work also clarifies why status concerns can generate wasteful expenditure and why status comparisons typically depend on the relevant audience and reference group (Frank, 1985; Hopkins and Kornienko, 2004).

A second literature investigates how status-driven consumption interacts with inequality, social stratification, and support for redistribution (Norman J Ireland, 1998; Hopkins, 2008; Hopkins and Kornienko, 2010; Gallice and Grillo, 2020). Several contributions show that status concerns can give high-income individuals an economic interest in redistribution. Levy and Razin (2015) show that when individuals engage in costly sorting (e.g., through schools or neighborhoods), redistribution can attract support even among above-mean voters because it reduces incentives to sort. Friedrichsen, König, and Lausen (2020) show that in dual provision systems richer citizens may favor higher public provision because it increases the exclusivity (status value) of private consumption. Adriani and Sonderegger (2019) show that redistribution can curb wasteful signaling, yet it may fail to secure majority support when voters value social distance from the poor. We contribute to this literature by introducing segmented status competition under partial class segregation and ordinal status concerns. In our framework, raising poor income can benefit the rich through a new channel: it intensifies poor–middle status competition in low-status environments, which shifts the middle class's status effort downward and relaxes status pressure at the top. The mechanism generates additional predictions (e.g., on charitable transfers and

poverty stigma) absent from existing frameworks, and it operates without changing the income ranking.

Our mechanism builds on an ordinal view of status, which naturally turns status into an object of competition: individuals care about being recognized as belonging to a higher social category and therefore choose signals to deter pooling and secure separation. This emphasis connects to work showing that equality can intensify wasteful status competition under ordinal concerns (Hopkins and Kornienko, 2004; Hopkins and Kornienko, 2006), while alternative (cardinal) formulations can deliver different comparative statics (Bilancini and Boncinelli, 2012; Hopkins, 2023). Relatedly, Hopkins and Kornienko (2010) distinguishes between equality in endowments and equality in the returns to status, which can have opposite effects on status expenditure. Finally, our modeling choice of partial segregation is motivated by both theory and evidence on reference groups. Some papers study status competition in networks in which each agent's neighborhood shapes comparisons (Ghiglino and Goyal, 2010; Immorlica et al., 2017), and Antinyan, Horváth, and Jia (2019) shows that when the network of status-concerned agents is endogenous, long-run links tend to form among individuals with similar income. Empirically, Mijs and Roe (2021) documents pervasive socioeconomic segregation across friendship networks, romantic partnerships, neighborhoods, education, workplaces, and labor markets, while Krivo et al. (2013) emphasizes that both disadvantaged and privileged districts can be socially isolated because daily activities occur in non-overlapping spaces. These patterns motivate our assumption that the extremes of the income distribution need not interact directly, whereas the middle class can interact with both. From a theoretical point of view, we abstract from settings where additional information about types can violate single-crossing and generate counter-signaling (Feltovich, R. Harbaugh, and To, 2002) or double-crossing preferences (C.-H. Chen, Ishida, and Suen, 2022). Instead, we obtain non-standard incentives through audience segmentation: different receivers observe different signals because social interactions occur in distinct environments, so non-monotonic incentives arise from segmented status competition rather than from high

types withholding signals.

1.3 Model

The present model is a variant of the status signaling model by Bilancini and Boncinelli (2012), inspired by L. S. Bagwell and Bernheim (1996), with two notable additions: a three-type population structure and partial segregation between social classes.

We model social status competition as a signaling environment with a continuum of senders and two audiences. A game is

$$\mathcal{G} = \langle T, (\alpha, \beta, \gamma), (I_t)_{t \in T}, (A_t)_{t \in T}, (u, s_\ell, s_h) \rangle,$$

and it is schematically illustrated in Figure 1.

Types. There is a unit mass of senders. Each sender's type is drawn independently from the common distribution over T with probabilities $\Pr(t = \mathcal{P}) = \alpha$, $\Pr(t = \mathcal{M}) = \beta$, $\Pr(t = \mathcal{R}) = \gamma$, where $\alpha + \beta + \gamma = 1$. The type t is privately observed by the sender and determines their income I_t, with $I_\mathcal{P} < I_\mathcal{M} < I_\mathcal{R}$. For tractability, we assume income levels are sufficiently separated so that all strategic considerations leave the income ranking unchanged.

Segregation and information. There are two locations $j \in \{\ell, h\}$[2]. In each location j, agents can make a locally observable conspicuous expenditure $x_j \geq 0$. Type $\mathcal{P}$ appears only in location ℓ, type $\mathcal{R}$ only in location h, and type $\mathcal{M}$ in both. Within each location, observers see only the local expenditure (receiver ℓ observes only x_ℓ and receiver h observes only x_h). No other characteristics, actions, or outcomes are observable. In particular, income and all inconspicuous consumption are never observed, and there is no information transmission across locations.

[2]I use 'location' and 'market' interchangeably to mean separate social arenas with distinct audiences and symbolic codes.

Actions. Action sets are:

$$A_{\mathcal{P}} = [0, I_{\mathcal{P}}], \qquad \text{choose } x_{\ell\mathcal{P}},$$

$$A_{\mathcal{M}} = \{(x_{\ell\mathcal{M}}, x_{h\mathcal{M}}) \in \mathbb{R}^2_+ : x_{\ell\mathcal{M}} + x_{h\mathcal{M}} \leq I_{\mathcal{M}}\}, \quad \text{choose } (x_{\ell\mathcal{M}}, x_{h\mathcal{M}}),$$

$$A_{\mathcal{R}} = [0, I_{\mathcal{R}}], \qquad \text{choose } x_{h\mathcal{R}}.$$

Given an action a_t, inconspicuous consumption is the residual income after conspicuous expenditure, namely:

$$I_{\mathcal{P}} - x_{\ell\mathcal{P}} \text{ for type } \mathcal{P}, \qquad I_{\mathcal{M}} - x_{\ell\mathcal{M}} - x_{h\mathcal{M}} \text{ for type } \mathcal{M}, \qquad I_{\mathcal{R}} - x_{h\mathcal{R}} \text{ for type } \mathcal{R}.$$

This residual is never publicly observed. Thus, only local conspicuous expenditure can affect beliefs and status.

Timing. (i) Nature draws t. (ii) The sender chooses $a_t \in A_t$. (iii) Receiver ℓ observes x_ℓ and receiver h observes x_h and form beliefs. (iv) Payoffs are realized.

Beliefs. Each location $j \in \{\ell, h\}$ hosts a passive audience (receiver) who observes only the local conspicuous expenditure x_j and forms beliefs about the sender's socioeconomic class. The audience takes no strategic action and affects payoffs only through the belief-dependent status term.

Receiver ℓ forms a belief about whether the sender is middle class, conditional on x_ℓ:

$$\mu_\ell(x_\ell) = \Pr(t = \mathcal{M} \mid x_\ell),$$

and receiver h forms a belief about whether the sender is rich, conditional on x_h:

$$\mu_h(x_h) = \Pr(t = \mathcal{R} \mid x_h).$$

Because income and the inconspicuous consumption are not publicly observed, only x_ℓ and x_h can affect beliefs.

Payoffs. The utility functions of the poor, the middle class, and the rich are formalized as follows.

$$U_{\mathcal{P}} = u(I_{\mathcal{P}} - x_{\ell\mathcal{P}}) + s_\ell(\mu_\ell(x_{\ell\mathcal{P}})),$$

$$U_{\mathcal{M}} = u(I_{\mathcal{M}} - x_{\ell\mathcal{M}} - x_{h\mathcal{M}}) + s_\ell(\mu_\ell(x_{\ell\mathcal{M}})) + s_h(\mu_h(x_{h\mathcal{M}})),$$

$$U_{\mathcal{R}} = u(I_{\mathcal{R}} - x_{h\mathcal{R}}) + s_h(\mu_h(x_{h\mathcal{R}})).$$

The first component, $u(\cdot)$, is utility from inconspicuous consumption (a composite private good, including savings), assumed strictly increasing and concave. The second component, $s_j : [0,1] \to \mathbb{R}_+$, represents the social status utility coming from being recognized by the audience in the visited location $j \in \ell, h$, as part of a given socio-economic class. The status-bearing object is the income, which, as well as inconspicuous consumption, is private information. The function $s_j(\cdot)$ is strictly increasing and continuous. We normalize ordinal status levels so that

$$s_\ell(0) = P, \quad s_\ell(1) = M, \quad s_h(0) = M, \quad s_h(1) = R,$$

which are assumed to be constant with respect to income changes and such that $P < M < R$.[3] When signaling is ineffective and types pool, agents' beliefs, and hence expected status payoffs, depend on the relative population shares (α, β, γ). By contrast, in separating equilibria, agents obtain full type-specific status, and equilibrium outcomes are independent of social composition.

People have ordinal concerns for social status, which means being concerned about their relative position in the socio-economic hierarchy, rather than the cardinal distance between ranks. We do not investigate here why concerns may be ordinal rather than cardinal (see Bilancini and

[3] Allowing P, M, R to depend on (α, β, γ) is a natural extension. Since the direction of this dependence is theoretically ambiguous and context-specific, we leave it outside the baseline model and interpret our results as holding for given status values P, M, R. The ambiguity arises because the social value of belonging to a given class can plausibly increase or decrease with its relative size. On the one hand, membership in a small group may carry greater exclusivity and therefore higher status (e.g., being rich in a society where the rich are a tiny elite). On the other hand, a larger group may command greater social recognition, visibility, or political salience, which can also raise the status associated with belonging to it (e.g., the middle class in societies where it constitutes the social "norm").

Boncinelli, 2014, for a discussion of the microfoundation of the shape of the status function).

Thanks to the strict concavity of $u(\cdot)$ and the ordering $I_{\mathcal{P}} < I_{\mathcal{M}} < I_{\mathcal{R}}$, the payoff function $u(I_t - x_j) + s_j(\mu_j)$ satisfies the Spence–Mirrlees condition in each location. Hence, indifference curves satisfy the single-crossing property in the (x_j, μ_j) space.

Strategies and equilibrium. A strategy profile is $\sigma = (\sigma_{\mathcal{P}}, \sigma_{\mathcal{M}}, \sigma_{\mathcal{R}})$ with $\sigma_t \in A_t$. An assessment $(\sigma, \mu_\ell, \mu_h)$ is a Weak Perfect Bayesian Equilibrium (WPBE) if: (i) each σ_t maximizes U_t given (μ_ℓ, μ_h); (ii) beliefs are consistent with σ via Bayes' rule on the equilibrium path.

The assumptions presented lead to two traditional signaling games bridged by the presence of the middle class in both. While the maximization problems of the poor and the rich are identical to those in conventional signaling games, and can therefore be analyzed separately, the middle class faces two interdependent signaling choices that necessitate a joint equilibrium analysis.

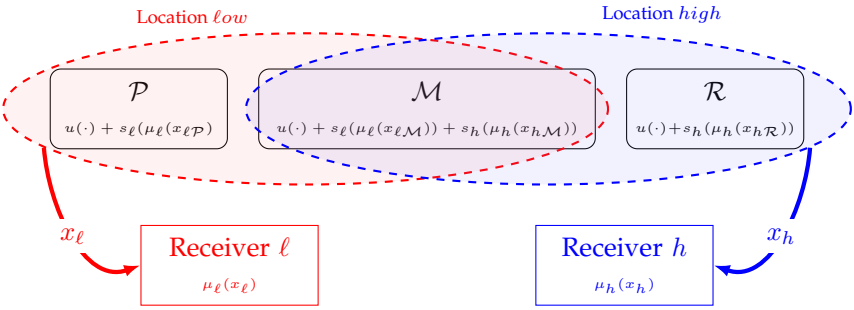


Figure 1: *Graphical representation of the model. The three black rounded boxes represent the social classes, each annotated with its corresponding utility. The red and blue shaded regions represent, respectively, the low and high location where individuals engage in conspicuous consumption, x_ℓ and x_h, to signal their status to the corresponding receivers (shown as colored square boxes), who in turn form beliefs $\mu_\ell(x_\ell)$ and $\mu_h(x_h)$.*

1.4 Equilibrium

This Section characterizes equilibrium behavior in a two-location status-signaling environment. We proceed in three steps. First, we introduce the equilibrium concept (WPBE) for this framework and show that the model admits multiple equilibrium classes. Second, we refine the equilibrium set by adapting the Intuitive Criterion to a two-location setting. Third, we derive a unique prediction at the level of outcomes (on-path expenditures) and explain intuitively why the refinement selects it.

We begin by defining the equilibrium concept and establishing multiplicity.

Definition 1 (WPBE). *An assessment* $(\hat{\sigma} = (\hat{x}_{\ell\mathcal{P}}, (\hat{x}_{\ell\mathcal{M}}, \hat{x}_{h\mathcal{M}})); \hat{x}_{h\mathcal{R}}; \mu)$*, with* $\mu = (\mu_\ell, \mu_h)$*, constitutes an overall Weak Perfect Bayesian Equilibrium in our three-type, two-location signaling game, if*

(i) $((\hat{x}_{\ell\mathcal{P}}, \hat{x}_{\ell\mathcal{M}}), \mu_\ell(x_\ell))$ *is WPBE in the signaling game played in location* ℓ*, whenever the middle-income type spends* $\hat{x}_{h\mathcal{M}}$ *in location* h*;*

(ii) $((\hat{x}_{h\mathcal{M}}, \hat{x}_{hR}), \mu_h(x_h))$ *is WPBE in the signaling game played in location* h*, whenever the middle-income type spends* $\hat{x}_{\ell\mathcal{M}}$ *in location* ℓ*;*

(iii) There is no cross-location profitable deviation for the middle-income type. Namely, for all $(x_{\ell\mathcal{M}}, x_{h\mathcal{M}})$ *such that* $x_{\ell\mathcal{M}} + x_{h\mathcal{M}} \leq I_{\mathcal{M}}$*,*

$$u(I_{\mathcal{M}} - \hat{x}_{\ell\mathcal{M}} - \hat{x}_{h\mathcal{M}}) + s_\ell(\mu_\ell(\hat{x}_{\ell\mathcal{M}})) + s_h(\mu_h(\hat{x}_{h\mathcal{M}})) \geq u(I_{\mathcal{M}} - x_{\ell\mathcal{M}} - x_{h\mathcal{M}}) + s_\ell(\mu_\ell(x_{\ell\mathcal{M}})) + s_h(\mu_h(x_{h\mathcal{M}})).$$

Definition 1 formalizes equilibrium in a way that respects the two-location structure: behavior must be locally consistent with signaling incentives in each location, while the middle type must have no profitable reallocation of conspicuous expenditure across locations.

The interaction of two local signaling problems, linked by the middle type's shared budget constraint, generates a multiplicity of equilibria. The next result summarizes existence across the four natural outcome configurations (pooling versus separating in each location).

Proposition 1 (Existence). *The game* $\mathcal{G}$ *admits at least one WPBE in each of the following classes: pooling–pooling, separating–separating, pooling–separating, and separating–pooling.*

Appendix 1.9.1 provides the proof of Proposition 1 and characterizes equilibria into the four classes.

Given this multiplicity, we next refine the equilibrium set by restricting off-path beliefs. As a technical note, because our environment involves two audiences and a two-dimensional choice for the middle type, standard refinements defined for canonical signaling games must be adapted. Definition 2 provides the corresponding two-location version used in the analysis.

Definition 2 (Two-location Intuitive Criterion)**.** *Fix a WPBE assessment (σ^*, μ^*) with equilibrium payoffs U_t^* for $t \in \{\mathcal{P}, \mathcal{M}, \mathcal{R}\}$. For any candidate deviation, define the set of types for whom the deviation can be made weakly profitable under some beliefs.*

***Deviations observed in location** ℓ. Given an off-path observation $x_\ell \geq 0$, define*

$$D_\ell(x_\ell) = \Big\{ t \in \{\mathcal{P}, \mathcal{M}\} : \exists\, \hat{\mu}_\ell \in [0,1] \text{ s.t. } \widetilde{U}_t(x_\ell; \hat{\mu}_\ell) \geq U_t^* \Big\},$$

where $\widetilde{U}_\mathcal{P}(x_\ell; \hat{\mu}_\ell) = u(I_\mathcal{P} - x_\ell) + s_\ell(\hat{\mu}_\ell)$, and for the middle type

$$\widetilde{U}_\mathcal{M}(x_\ell; \hat{\mu}_\ell) = \max_{\substack{x_h \geq 0 \\ x_\ell + x_h \leq I_\mathcal{M}}} \Big\{ u(I_\mathcal{M} - x_\ell - x_h) + s_\ell(\hat{\mu}_\ell) + s_h(\mu_h^*(x_h)) \Big\}.$$

That is, after deviating in ℓ, the middle type is allowed to reallocate conspicuous expenditure across locations, anticipating that the audience in h updates according to μ_h^* upon observing x_h.

***Deviations observed in location** h. Given an off-path observation $x_h \geq 0$, define*

$$D_h(x_h) = \Big\{ t \in \{\mathcal{M}, \mathcal{R}\} : \exists\, \hat{\mu}_h \in [0,1] \text{ s.t. } \widetilde{U}_t(x_h; \hat{\mu}_h) \geq U_t^* \Big\},$$

where $\widetilde{U}_\mathcal{R}(x_h; \hat{\mu}_h) = u(I_\mathcal{R} - x_h) + s_h(\hat{\mu}_h)$, and

$$\widetilde{U}_\mathcal{M}(x_h; \hat{\mu}_h) = \max_{\substack{x_\ell \geq 0 \\ x_\ell + x_h \leq I_\mathcal{M}}} \Big\{ u(I_\mathcal{M} - x_\ell - x_h) + s_\ell(\mu_\ell^*(x_\ell)) + s_h(\hat{\mu}_h) \Big\}.$$

Restriction on off-path beliefs. *The WPBE (σ^*, μ^*) satisfies the Intuitive Criterion if, for every off-path x_ℓ and x_h:*

- *if* $D_\ell(x_\ell) = \{\mathcal{P}\}$ *then* $\mu^*_\ell(x_\ell) = 0$, *and if* $D_\ell(x_\ell) = \{\mathcal{M}\}$ *then* $\mu^*_\ell(x_\ell) = 1$;
- *if* $D_h(x_h) = \{\mathcal{M}\}$ *then* $\mu^*_h(x_h) = 0$, *and if* $D_h(x_h) = \{\mathcal{R}\}$ *then* $\mu^*_h(x_h) = 1$;

while if both types in the location j *belong to* $D_j(\cdot)$, *the criterion imposes no restriction.*

In words, the criterion restricts off-path beliefs by requiring that, after an unexpected expenditure in a given location, the audience should not assign positive probability to a type for whom the deviation cannot be rationalized as weakly profitable under any belief. Crucially, the middle type is allowed to reallocate expenditure across locations when evaluating deviations.

We now turn from equilibrium *beliefs* to equilibrium *behavior*. While WPBE may differ in how they specify off-path beliefs, the two-location Intuitive Criterion pins down a unique pattern of *on-path* conspicuous expenditures. To state this prediction concisely, we introduce three expenditure cutoff levels. These cutoffs summarize the key incentive constraints: separation in location ℓ (1.1), separation in location h (1.2), and the middle type's incentive to reallocate expenditure across locations (1.3).

Definition 3 (Cutoff expenditures)**.** *Define* $x^*_\ell > 0$ *as the unique solution to the poor type's indifference condition in location* ℓ*:*

$$u(I_\mathcal{P}) + P = u(I_\mathcal{P} - x^*_\ell) + M. \tag{1.1}$$

Given x^*_ℓ, *define* $x^*_h > 0$ *as the (Riley) separating expenditure in location* h *that deters type* $\mathcal{M}$ *from mimicking type* $\mathcal{R}$ *while holding fixed* x^*_ℓ*:*

$$u(I_\mathcal{M} - x^*_\ell) + M = u(I_\mathcal{M} - x^*_\ell - x^*_h) + R. \tag{1.2}$$

Finally, define $x^{**}_h \geq 0$ *as the minimal expenditure that deters the middle type's* cross-location *deviation, allowing it to reallocate conspicuous spending across locations:*

$$u(I_\mathcal{M} - x^*_\ell) + 2M = u(I_\mathcal{M} - x^{**}_h) + P + R. \tag{1.3}$$

Let

$$\tilde{x}_h \equiv \max\{x^*_h, x^{**}_h\}. \tag{1.4}$$

The term x_h^* corresponds to the standard Riley separating expenditure in location h absent cross-location reallocations, while x_h^{**} captures the additional deterrence needed when the middle type can shift expenditure across locations; $\tilde{x}_h$ is the effective separating level selected in equilibrium.

The economic interpretation of the three cutoffs is straightforward. The cutoff x_ℓ^* is the minimal expenditure the middle type must incur in location ℓ to avoid being mistaken for the poor. The cutoff x_h^* is the standard within-location Riley expenditure needed for the rich to deter direct imitation by the middle in location h. By contrast, x_h^{**} captures a distinctly two-location force: it is the expenditure needed to deter the middle type from reallocating conspicuous spending away from ℓ and toward h in order to seek recognition as rich. Thus, while x_h^* reflects a local signaling constraint, x_h^{**} reflects the fact that the middle class links the two status arenas through its shared budget.

These cutoffs allow us to express the refined equilibrium outcome in closed form.

Proposition 2 (Unique outcome)**.** *The two-location Intuitive Criterion selects a unique equilibrium outcome (up to off-path beliefs). On the equilibrium path:*

$$x_{\ell\mathcal{P}} = 0, \qquad x_{\ell\mathcal{M}} = x_\ell^*, \qquad x_{h\mathcal{M}} = 0, \qquad x_{h\mathcal{R}} = \tilde{x}_h.$$

Equivalently, the selected outcome is: (i) Riley–Riley *if* $x_h^* \geq x_h^{**}$*; and (ii)* Riley–Non-Riley *(with rich overspending in location h) if* $x_h^{**} > x_h^*$.

The proof of Proposition 2 is provided in Appendix 1.9.3. Detailed definitions of the Riley–Riley and Riley–Non-Riley configurations are presented in Appendix 1.9.2. In the Riley–Riley outcome, the rich separate by spending in h just enough to deter within-h imitation by the middle, while the middle spends in ℓ to avoid being perceived as poor. In the Riley–Non-Riley outcome, the rich must also deter the middle's cross-location reallocation (shifting expenditure from ℓ to h), so equilibrium spending in h can exceed the within-h Riley level.

The key economic force is whether the rich only need to discourage imitation within location h, or must also discourage the middle type from abandoning downward separation in location ℓ in order to compete more

aggressively upward. In the Riley–Riley case, the standard within-h separating expenditure is already sufficient: once the middle type pays x_ℓ^* to distinguish itself from the poor, it is too costly to redirect enough spending toward h to mimic the rich. In the Riley–Non-Riley case, instead, the middle type has a sufficiently strong incentive to seek recognition at the top that the rich must overspend relative to the standard Riley benchmark. This overspending arises entirely from the two-location structure: because the middle class operates in both arenas, status pressure at the bottom affects the intensity of competition at the top.

Proposition 2 summarizes the model's unique behavioral prediction under the refinement. The two-location Intuitive Criterion eliminates all pooling outcomes. Pooling requires off-path beliefs under which a deviation to a more status-intensive expenditure is interpreted as coming from a lower type, even though only higher types can profitably deviate. As in standard signaling models, such beliefs are not credible: whenever a deviation can benefit only the higher type, the audience should assign it to that type. Hence, pooling cannot survive in either location.

Among the remaining separating outcomes, the criterion selects the most efficient one, i.e. the separating allocation that achieves incentive compatibility at the lowest wasteful conspicuous spending. In location ℓ, efficiency implies a Riley outcome: the poor spend zero, and the middle class spends the minimal amount x_ℓ^* that deters imitation. In location h, the rich must separate efficiently from the middle class. If the standard Riley level x_h^* already prevents the middle class from reallocating expenditure away from ℓ to mimic the rich in h, the selected equilibrium has a Riley–Riley configuration. Otherwise, the rich must increase spending to x_h^{**} to block this cross-location deviation, yielding a Riley–Non-Riley configuration.

Which configuration arises depends on the model parameters, in particular on how attractive rich recognition is for the middle type relative to the utility cost of additional signaling and the status loss from being perceived as poor. When the status value of being recognized as rich is relatively low, or when being perceived as poor is sufficiently costly, the middle type has little incentive to reallocate expenditure upward, so

the standard within-location separating constraint in h is sufficient and the equilibrium is Riley–Riley. By contrast, when recognition as rich becomes sufficiently valuable, the middle type is more strongly tempted to shift expenditure from ℓ to h in order to compete upward. In that case, the cross-location incentive constraint becomes binding, and the equilibrium switches to Riley–Non-Riley. Appendix 1.10 illustrates this mechanism by showing how the selected configuration varies with the status utility from being recognized as rich, holding all other parameters constant.

1.5 Segregation

The introduction of segregation, modeled as two separated locations in which different classes compete for status, has non-trivial implications for wasteful conspicuous spending. In this Section, we compare the equilibrium predictions of our model to a benchmark environment in which all classes compete in a single location. We show that segregation benefits the rich, which in turn suggests an incentive for the upper classes to favor social arrangements that increase separation from lower-income groups. More specifically, we expect the wealthy to support the emergence of new forms of conspicuous consumption that reinforce social boundaries and sustain exclusionary status hierarchies.

Benchmark (one location). In the benchmark case, the three classes compete for social status within the same location, purchasing a single conspicuous good x. By a standard result in signaling models, the equilibrium features a separating (Riley) outcome such that the equilibrium levels of conspicuous expenditure of the poor, middle, and rich are respectively

$$x_{\mathcal{P}}^{one} = 0, \qquad x_{\mathcal{M}}^{one} > 0, \qquad x_{\mathcal{R}}^{one} > 0,$$

where $x_{\mathcal{M}}^{one}$ and $x_{\mathcal{R}}^{one}$ satisfy

$$u(I_{\mathcal{P}}) + P \;=\; u(I_{\mathcal{P}} - x_{\mathcal{M}}^{one}) + M, \tag{1.5}$$

$$u(I_{\mathcal{M}} - x_{\mathcal{M}}^{one}) + M \;=\; u(I_{\mathcal{M}} - x_{\mathcal{R}}^{one}) + R. \tag{1.6}$$

Figure 2 depicts this benchmark outcome.

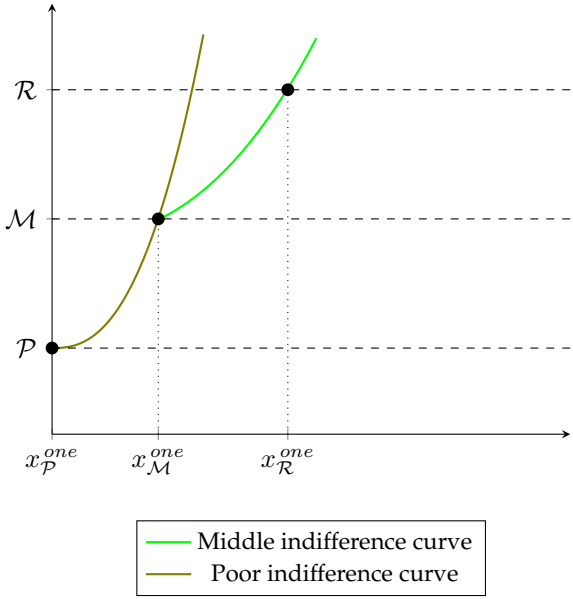


Figure 2: *Riley outcome of a three-type, one-location signaling game, representing status competition in the absence of segregation. Equilibrium conspicuous spending is* $x_{\mathcal{P}}^{one} = 0$, $x_{\mathcal{M}}^{one}$, *and* $x_{\mathcal{R}}^{one}$.

Partial segregation via time allocation. To assess the role of segregation while keeping the total amount of status at stake comparable to the benchmark, we extend the two-location model by assuming that the middle class spends an exogenous fraction $\tau \in (0, 1)$ of time in location ℓ and the remaining fraction $(1-\tau)$ in location h. Accordingly, the middle type evaluates status as a convex combination of the status obtained in the two locations. The parameter τ is exogenous and cannot be chosen by any type.

We restrict attention to values of τ for which the single-crossing property continues to hold in the relevant comparisons in each location.[4] Under this restriction, the expenditures separating $\mathcal{P}$ from $\mathcal{M}$ in location ℓ are unchanged: $x_{\ell\mathcal{P}} = 0$ and $x_{\ell\mathcal{M}} = x_{\ell}^{*}$ as defined in (1.1).

[4]If τ is close to 0, the middle class places negligible weight on status in location ℓ and may prefer not to signal there, potentially breaking the single-crossing logic.

In location h, define $x_h^*(\tau) > 0$ as the (Riley) separating expenditure that deters within-location imitation by $\mathcal{M}$ holding x_ℓ^* fixed,

$$u(I_{\mathcal{M}} - x_\ell^*) + (1-\tau)M = u(I_{\mathcal{M}} - x_\ell^* - x_h^*(\tau)) + (1-\tau)R, \tag{1.7}$$

and define $x_h^{**}(\tau) \geq 0$ as the minimal expenditure that deters the middle type's cross-location reallocation deviation,

$$u(I_{\mathcal{M}} - x_\ell^*) + M = u(I_{\mathcal{M}} - x_h^{**}(\tau)) + \tau P + (1-\tau)R. \tag{1.8}$$

The rich must deter both deviations, hence the equilibrium expenditure in h is

$$\tilde{x}_h(\tau) \equiv \max\{x_h^*(\tau), x_h^{**}(\tau)\}. \tag{1.9}$$

Segregation reduces rich spending and raises rich payoffs. The next lemma shows that segregation alleviates the rich type's separation problem and consequently reduces the rich type's equilibrium conspicuous expenditure relative to the one-location benchmark.

Lemma 1 (Segregation lowers the rich type's equilibrium expenditure)**.** *Let $\tau \in (0,1)$, let $x_{\mathcal{R}}^{one}$ solve* (1.6)*, and let $\tilde{x}_h(\tau)$ solve* (1.9)*. Then*

$$\tilde{x}_h(\tau) = \max\{x_h^*(\tau), x_h^{**}(\tau)\} < x_{\mathcal{R}}^{one}.$$

Proof of Lemma 1 is in Appendix 1.9.4.

When competition is split across two locations, mimicking the rich in h no longer distinguishes the middle from the poor "for free," because the middle must still spend in location ℓ to avoid being perceived as poor there. This opportunity cost weakens the middle as a competitor in h, allowing the rich to separate with lower expenditure. Moreover, the middle values status in h only for a fraction $(1-\tau)$ of time, which further reduces the incentive to mimic. Since the rich remains recognized as rich on the equilibrium path, lower conspicuous expenditure translates into higher consumption and therefore higher utility.

In both environments, the rich type attains the top rank on the equilibrium path (it is recognized as rich), so the comparison is driven by the consumption component. Since $u(\cdot)$ is strictly increasing, Lemma 1 implies Result 1.

Result 1 (Segregation is profitable for the rich)**.** *Under the assumptions of Lemma 1, segregation increases the rich type's equilibrium payoff.*

A further observation regarding segregation is that, as the middle class spends more time in location ℓ (higher τ), its incentive to compete for high-status recognition in h weakens. This relaxation of status competition loosens the rich type's separation constraint in h and allows the rich to save on conspicuous spending. Put differently, segregation operates as a substitute for expenditure: by splitting status competition across locations, a higher τ reduces the conspicuous signal required for the rich to maintain separation, sustaining the same status distance with less consumption.

1.6 Income distribution

Status concerns shape how people evaluate redistributive policies. We examine how a marginal increase $\delta > 0$ in the poor's income reshapes equilibrium conspicuous spending ($x_\ell^{*\delta}$ and $\tilde{x}_h^\delta = \max\{x_h^{*\delta}, x_h^{**\delta}\}$) and the rich's incentives to support redistribution in an economy with partial segregation.

A higher income of the poor intensifies status competition in location ℓ. Intuitively, when the poor become relatively richer, the middle class must increase conspicuous spending in ℓ to preserve separation. The left panel of Figure 3 illustrates this mechanism.

The same shock affects incentives in the high-status location h because it changes the middle class's outside options and constraints. The key trade-off is two-sided. On the one hand, since the middle class spends more in ℓ, it has less disposable income available to compete for status in h. On the other hand, tougher competition in ℓ makes it more attractive for the middle class to attempt a cross-location deviation toward h. Which force dominates depends on the configuration of incentive constraints that the rich satisfy in equilibrium.

When the equilibrium has a *Riley–Riley* structure, the rich need to deter within-location imitation: they choose their conspicuous spending so as to make the middle class indifferent between competing for the rich's status in h and not doing so. In this case, a higher $I_{\mathcal{P}}$ relaxes the relevant constraint in h because the middle class becomes effectively "poorer" in h after spending more in ℓ. The right panel of Figure 3 depicts this case.

When the equilibrium instead has a *Riley–Non-Riley* structure, the rich need to deter *cross-location* deviations by the middle class. Then, making ℓ more competitive strengthens the middle class's incentive to deviate across locations, and the rich must increase conspicuous spending in h to keep such deviations unattractive.

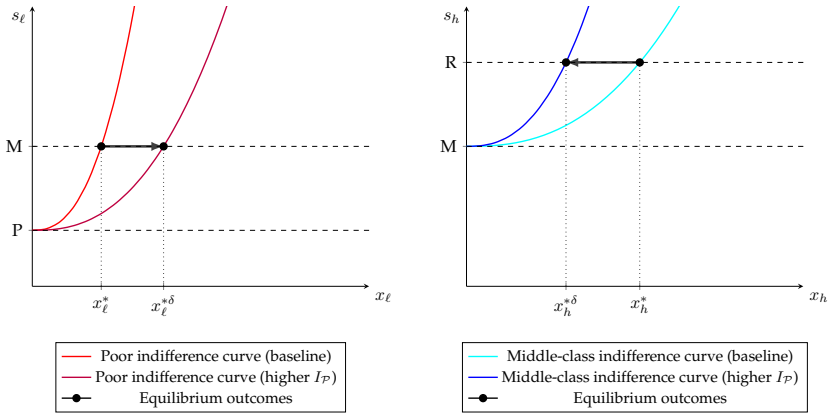


Figure 3: ***Poor's income increase*** *Left panel: a higher income of the poor flattens the poor's indifference condition, raising the middle class's equilibrium expenditure from x_ℓ^* to $x_\ell^{*\delta}$. Right panel: in the Riley–Riley configuration, the same shock steepens the middle class's relevant indifference condition in h, allowing the rich to reduce equilibrium expenditure from x_h^* to $x_h^{*\delta}$.*

Proposition 3 summarizes the comparative statics: increasing the income of the poor always tightens competition in ℓ, but it helps the rich in the Riley–Riley configuration (because it relaxes within-h imitation incentives) and hurts them in the Riley–Non-Riley configuration (because it strengthens cross-location deviation incentives).

Proposition 3 (Mechanism)**.** *A marginal increase in the poor's income affects the rich's equilibrium conspicuous spending in location h, depending on the equilibrium configuration:*

(i) Riley–Riley. *A marginal increase in the income of the poor reduces the rich's equilibrium conspicuous spending in h, so the rich are better off.*

(ii) Riley–Non-Riley. *A marginal increase in the income of the poor increases the rich's equilibrium conspicuous spending in h, so the rich are worse off.*

Appendix 1.9.5 provides the proof of Proposition 3.

In what follows, we focus on the Riley–Riley configuration, because it creates scope for policy intervention: raising $I_\mathcal{P}$ generates a status-based surplus for the rich by reducing wasteful expenditures in h.

This naturally raises the question of how such targeted resources can be financed. A natural benchmark is voluntary giving by the rich. Corollary 1 formalizes this intuition by showing that in the Riley–Riley configuration, the rich are willing to transfer resources to the poor if and only if the status-based benefit outweighs the direct monetary cost of the transfer.

Corollary 1 (When do the rich donate?)**.** *In the Riley–Riley case, the rich support a positive transfer when this status-based gain offsets the direct monetary loss, namely*

$$\frac{\gamma}{\alpha} \geq \frac{1}{\left(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x_\ell^*)}\right)\left(1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}\right)}. \tag{1.10}$$

Appendix 1.9.6 provides the proof of Corollary 1.

Condition (1.10) highlights a simple aggregate feasibility logic: a given donation by the rich is diluted across the poor, so the per-poor increase in $I_{\mathcal{P}}$ scales with γ/α. As a result, voluntary donations become rational only when the mass of rich is sufficiently large relative to the mass of poor. Specifically, concavity implies that each bracket in the denominator lies in $(0, 1)$, the right-hand side exceeds 1, so a necessary condition is $\gamma/\alpha > 1$. This requirement can be demanding: even though the status channel is positive in the Riley–Riley configuration, a small rich group cannot generate enough per-poor income gains to justify the private cost.

Because voluntary transfers rely on this strong demographic condition, we next consider a policy that taxes the rich and the middle class and redistributes the revenues among the poor. Corollary 2 characterizes when such a policy can be sustained by majority voting.

Corollary 2 (When does a lump-sum tax on $\mathcal{M}$ and $\mathcal{R}$ win a majority?)**.** *In the Riley–Riley case, consider a balanced-budget policy indexed by $\kappa > 0$ that taxes each middle-class and rich individual a lump-sum amount κ and rebates the proceeds equally to the poor. Thus, each poor individual receives the subsidy*

$$\Delta I_{\mathcal{P}}(\kappa) = \frac{1 - \alpha}{\alpha}\,\kappa.$$

Then there exists $\bar{\kappa} > 0$ such that a policy with any $\kappa \in (0, \bar{\kappa}]$ is supported by the majority of the population if

$$\left(1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}\right)\left(1 + \frac{1-\alpha}{\alpha}\Big(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x_\ell^*)}\Big)\right) \geq 1. \quad (1.11)$$

$$\textit{and} \qquad \alpha + \gamma \geq \frac{1}{2}. \quad (1.12)$$

Appendix 1.9.7 provides the proof of Corollary 2.

Relative to Corollary 1, the political condition is more permissive: the rich's support requires only that the poor are not "too many" (1.11). In that case, the poor and the rich together can form a coalition that constitutes a majority under condition (1.12). This captures the idea that the rich's status-based gain can sustain an interclass coalition with the poor even when voluntary donations are not individually rational. At the same time, the middle class bears a double burden in this policy: it pays the tax and, because $I_{\mathcal{P}}$ rises, it must also spend more in ℓ to preserve separation.
Result 2 summarizes the message of this Section.

Result 2 (Redistribution and interclass coalitions under segregation)**.** *With partial segregation, status competition can make donations to the poor attractive to the rich and can sustain rich–poor political coalitions, provided the type distribution allows a winning majority and the equilibrium has a Riley–Riley structure.*

Both corollaries can be translated in terms of progressive income taxation with lump-sum redistribution. The first corollary corresponds to a setting in which the rich effectively finance targeted increases in $I_{\mathcal{P}}$ (leaving the middle class's budget unaffected directly), while the second allows the tax burden to fall on both the rich and the middle class and therefore admits the possibility of a negative net effect on the middle class. This distinction matters for implementation. If a policymaker evaluates redistribution only through direct monetary transfers and ignores status-driven responses, she may view a progressive lump-sum redistribution as broadly acceptable, while in equilibrium it can impose

additional (and politically salient) costs on the middle class through intensified status competition in ℓ. In this sense, status concerns and segregation shape not only whether redistribution is desirable for the rich, but also which redistributive instruments are politically sustainable once we account for endogenous conspicuous spending.

1.7 Poverty stigma

Redistribution is generally intended to reduce inequality and ensure access to basic necessities, including food, housing, and healthcare. However, redistributive policies are often accompanied by social stigma: individuals receiving transfers may be perceived as lazy or undeserving, which lowers their social standing (Baumberg, 2016). In this Section, we show that poverty stigma systematically benefits the rich by reshaping status competition.

We model stigma as a parameter $B > 0$ that reduces the social status associated with being recognized as poor. That is, after the policy, the poor obtain status $P - B$ rather than P. Lowering the poverty status intensifies competition in location ℓ, since the poor are willing to spend more to avoid stigma, and the middle class must respond by increasing its separating expenditure. This mechanism, formalized in Proposition 4, relaxes the separation constraint faced by the rich and reduces their equilibrium conspicuous spending.

Proposition 4 (Poverty stigma benefits the rich)**.** *Introducing poverty stigma ($B > 0$) reduces the equilibrium conspicuous spending of the rich in both the Riley–Riley and the Riley–Non-Riley configurations.*

The proof of Proposition 4 is in Appendix 1.9.8.

Since poverty stigma provides the rich with an additional way to reduce their expenditures, we next examine the conditions under which a transfer that stigmatizes poverty can be implemented. We do so in Corollary 3.

Corollary 3 (When is a stigmatizing transfer jointly supported?)**.** *Consider a redistributive policy that transfers a fraction t of the rich's income to the poor and induces poverty stigma of magnitude $B > 0$.*

(i) The poor accept the policy if and only if the monetary gain outweighs the stigma cost,

$$B < u\Big(I_{\mathcal{P}} + \frac{\gamma}{\alpha} t I_{\mathcal{R}}\Big) - u(I_{\mathcal{P}}). \tag{1.13}$$

(ii) In a Riley–Riley configuration, the rich support the policy if and only if stigma is sufficiently strong to offset the direct monetary cost of the

transfer,

$$B > tI_{\mathcal{R}} \left[\frac{u'(I_{\mathcal{P}} - x_\ell^*)}{1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}} - \frac{\gamma}{\alpha} \big(u'(I_{\mathcal{P}} - x_\ell^*) - u'(I_{\mathcal{P}}) \big) \right]. \quad (1.14)$$

A stigmatizing transfer is jointly supported by the rich and the poor if and only if the interval defined by (1.13) *and* (1.14) *is non-empty, which requires*

$$\frac{\gamma}{\alpha} > \frac{1}{1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}}. \quad (1.15)$$

The proof of Corollary 3 is in Appendix 1.9.9.

Under (1.14), the rich are willing to transfer a fraction t of their income to the poor. While sufficiently strong stigma makes the transfer attractive to the rich, it simultaneously reduces the poor's willingness to accept redistribution. Joint support, therefore, requires an intermediate range of stigma levels, which exists only when condition (1.15) holds. In particular, a necessary condition for political feasibility is that the rich outnumber the poor.

In our framework, poverty stigma is modeled in reduced form as an exogenous parameter $B > 0$ that lowers the social status associated with being identified as poor. This choice is mainly analytical: it allows us to isolate the equilibrium effects of stigma on status competition and redistributive support without modeling the formation of stigma itself. At a deeper level, however, B can be interpreted as the outcome of social narratives, institutional design, and collective beliefs. Policies that publicly distinguish recipients from non-recipients, residential segregation, or discourses emphasizing deservingness may increase B, whereas universalistic programs, integration policies, and changes in public narratives may reduce it over time. In this sense, the model does not treat stigma as normatively or historically fixed; rather, it studies how a given level of stigma reshapes equilibrium behavior, while leaving its social production and persistence to future work.

This mechanism suggests a possible explanation for why redistributive programs that impose stigma on recipients may nonetheless attract

political support from higher-income groups, even though they harm beneficiaries directly, while non-stigmatizing universal transfers often encounter resistance. More broadly, the model highlights how stigma can function not merely as a social byproduct of redistribution but as a policy instrument that reshapes status competition and the formation of interclass coalitions. Result 3 summarizes the role of stigma for poverty.

Result 3 (Poverty stigma and redistributive support)**.** *Introducing poverty stigma reduces equilibrium conspicuous spending by the rich in both the Riley–Riley and Riley–Non-Riley configurations, by intensifying status competition between the poor and the middle class. As a result, stigma relaxes the incentive constraints faced by the rich and can expand the set of redistributive policies they are willing to support, even when such policies impose welfare losses on recipients.*

In contrast to the wealthy, the middle class has an interest in opposing the stigma of poverty. While stigma increases the symbolic distance between the poor and the middle class, it simultaneously intensifies competition in location ℓ. As a result, the middle class must raise its conspicuous spending to avoid being associated with the stigmatized poor. Because this additional expenditure is purely wasteful and does not improve the middle class's position with respect to the rich, stigma strictly reduces middle-class welfare. In the model, stigmatizing redistribution therefore places the middle class in a disadvantageous position: it raises their cost of maintaining status without generating compensating gains, explaining why the middle class has no incentive to support stigmatization.

More generally, the model sheds light on how different social classes may hold conflicting interests with respect to manipulations of the status distribution. In broad terms, the rich benefit from policies that lower the relative status of the middle class while easing pressure at the bottom; the middle class favors policies that reduce the status advantage of the rich while improving the position of the poor; and the poor benefit from policies that relax competition with the middle class, even if this implies a relative improvement in the status of the rich. Remark 1 provides a discipline to these conjectures by highlighting the constraints imposed

by a possible status normalization. A systematic analysis of the relative political power of different classes, however, lies beyond the scope of this paper. We therefore focus on poverty stigmatization, which emerges from the model as an empirically relevant prediction for which there is no formal explanation.

Remark 1 (Status normalization)**.** *Throughout the analysis, we only require $P < M < R$. A possible extension is that total social status is fixed and normalized so that $P + M + R = 1$. Under this normalization, any stigma-induced decrease in P must be offset by an increase in M and/or R. This reallocation does not automatically help the middle class. If M rises, middle-class status becomes more valuable, which can* strengthen *incentives to separate from the poor in location ℓ and therefore intensify status competition. In this sense, status normalization highlights a tension: stigma can raise the social stakes of being middle class while also increasing the spending needed to sustain middle-class distinctiveness.*

1.8 Discussion

In a signaling model of conspicuous consumption, high-income individuals have an economic incentive to promote social segregation. Segregation alters equilibrium signaling incentives by reducing the need for costly status displays among the rich, which in turn affects their preferences over redistribution. These forces contribute to the emergence and persistence of poverty stigma. The model's theoretical predictions help rationalize patterns observed in real-world societies.

Result 1 shows that segregation functions as a substitute for conspicuous consumption. This mechanism helps explain the tendency of high–income individuals to separate themselves from the rest of society: as discussed in Section 1.5, greater social segregation reduces the need for costly displays of status. The model's notion of spatial distance admits a broader interpretation. Returning to the example of language introduced earlier, affluent individuals have incentives not only to physically distance themselves from lower–income groups, but also to develop new, exclusive forms of signaling—"languages", that are inaccessible to others. Consistent with this view, evidence indicates that high–income households increasingly signal status through less visible channels, particularly investments in education and cultural capital. For example, Currid-Halkett (2019) documents that in the United States, the top 1% more than tripled the share of income devoted to education between 1996 and 2014, while education spending among middle–income households remained roughly constant.

Result 2 concerns redistribution. The relationship between support for redistribution and income is not uniform across contexts. Instead, cross-country evidence shows substantial heterogeneity in both the sign and strength of the income–redistribution gradient (Steele, Cohen, and van der Naald, 2022), pointing to mechanisms beyond simple income-based self-interest or linear altruism. We show that, under certain conditions, it is optimal for the rich to transfer resources to the poor, either through direct donations or by supporting redistributive policies. We do not presume that this mechanism fully explains the charitable actions

of wealthy people. However, it can be combined with altruism (Fong, 2001), biased views about the determinants of inequality (Bénabou and Tirole, 2006), prospects of social mobility (Piketty, 1995), and fear of social upheavals (Simmel and Jacobson, 1965) to better understand why people may support redistribution. Beyond this general prediction, the model delivers comparative statics: depending on the intensity of status concerns, the relative size of social classes, and the income distribution, different socioeconomic groups support different forms of redistribution.

In a segregated society where the middle class is obsessed with not looking poor (a Riley–Riley outcome), the equilibrium predicts that the rich purchase just enough of the *exclusive* conspicuous good to deter middle-class emulation, confident that the middle class will keep spending on the *common* conspicuous good to avoid being mistaken for poor. A realistic interpretation is that the common conspicuous good consists of widely recognizable status markers (branded clothes, visible accessories, and other "logo" goods) that the middle class uses to signal distance from poverty. By contrast, the exclusive conspicuous good captures forms of distinction that require both money and cultural competence to be recognized (a tailor-made suit with subtle cues, original works of art, or niche tastes and aesthetic codes) that are largely unintelligible to the lower part of society.

When the poor become richer, the middle class must increase spending on the common good to preserve separation from below. This intensified "downward" pressure relaxes the rich's incentive constraint and reduces the need to signal at the top. As a result, redistribution can become attractive to the rich: by raising the poor's income, transfers shift part of the status competition to the lower segment of society and reduce pressure on the rich.

From a policy perspective, if the middle class does not constitute a majority, such that the poor and the rich together form the larger group, formal policies transferring public resources to the poor are likely to be implemented. When the middle class constitutes the majority, instead, it can block such policies. Importantly, this opposition does not reflect hostility toward the poor *per se*, but rather the middle class's incentive

to protect its separating position. Finally, when the poor and the rich together represent a small fraction of the population and the rich outnumber the poor, the rich may prefer voluntary transfers: targeted donations to the poorest can generate sizable status gains while limiting the extent to which redistribution compresses the middle class's separating standard.

Conversely, in a segregated society where the middle class is dazzled by the social status of the rich (a Riley–Non-Riley outcome), the equilibrium predicts that the rich purchase just enough of the *exclusive* conspicuous good to prevent the middle class from cutting back on the *common* conspicuous good. In this way, the middle class is willing to accept the risk of being mistaken for poor by part of the broader public if doing so allows them to be perceived as rich within the elite social circles they occasionally access. In this case, redistribution harms the rich not only directly, through reduced income, but also indirectly through the status channel. Increasing the income of the poor intensifies status rivalry from below, making competition at the top relatively more attractive and forcing the rich to increase their conspicuous spending. Consequently, we expect to observe less generosity and fewer poverty-oriented redistributive policies in this type of society.

To grasp the relevance of this result, we need to gauge the relative size of income classes in the population. This is not straightforward, because the literature offers several competing ways to define the "middle class" (with the poor and the rich defined residually). Researchers typically use income-based definitions because income data are relatively comparable across countries and over time. They then operationalize "middle income" in three main ways: (i) absolute cutoffs, such as the World Bank's 10–50 per day in PPP terms (particularly useful for emerging economies) (Ferreira et al., 2012); (ii) a fixed segment of the income distribution, such as the "middle 60%," which is simple but mechanically prevents the middle class from shrinking or expanding by construction (Reeves, Guyot, and Krause, 2018; Pew Research Center, 2015); and (iii) relative thresholds around the median, for instance income between 75% and 200% of the national median, which allows the middle

class size to vary across countries and over time and is consistent with standard measures of relative poverty (OECD, 2019). Using this median-based definition, the middle-income group remains the largest in most advanced economies, yet its size has declined across generations. On average across OECD countries, the share of individuals in middle-income households (75%–200% of national median disposable income) fell from 64% to 61% between the mid-1980s and the mid-2010s (OECD, 2019). The United States illustrates how far this shift can go: Pew Research Center (2015) reports that the middle-income share declined from 61% in 1971 to 50% in 2015, implying that adults above and below the middle combined are now roughly as numerous as those in the middle. In these terms, the hypothesis that rich and poor may jointly form a majority, and thus potentially support policies at the expense of the middle class, becomes empirically plausible. Finally, the relevant "rich–poor coalition" in our model depends on who benefits from redistribution. In the data, some widely used income-class definitions classify as "middle income" households that weakly benefit from redistribution, suggesting that empirical "middle class" measures may include a portion of what our model labels as poor (Causa and Hermansen, 2019).

The partial segregation formalized in this model is essential to generate all the effects presented. In the absence of segregation, the potential status-based interest of the rich in increasing the income of the poor disappears. Thus, conditional on a Riley–Riley outcome, more integrated societies are expected to exhibit lower levels of charity and redistribution, *ceteris paribus*.

Result 3 shows that the rich can benefit from strengthening narratives that stigmatize poverty, i.e., to lower the social status attached to being poor. The role of poverty stigmatization studied in Section 1.7 is consistent with the evidence presented by C. Graham and Grisard (2019). As already emphasized by Simmel and Jacobson (1965), the poor may generate social and economic benefits for the wealthy by serving as recipients of aid. Graham and colleagues document that, in nineteenth-century Canada, charitable organizations often subjected recipients to humiliating procedures, which they interpret as a transfer of social status from

	Riley–Riley	Riley–Non-Riley
Segregation	The rich benefit from partial social segregation.	
Transfer to the poor	The rich support a transfer, but are willing to make direct donations only if they outnumber the poor.	The rich oppose a transfer.
Poverty stigma	The rich benefit from stigmatizing the poor.	

Table 1: Summary of the results

the poor to the rich. In our model, the rich have an additional indirect interest in stigmatizing poverty, which provides an additional explanation for these findings based on the intensified competition faced by the middle class.

Another relevant pattern of social status recognition concerns modern society's tendency to celebrate the middle class. A prominent example is the image of the white picket fence, an iconic symbol of the idealized middle-class suburban lifestyle, representing family stability, home ownership, and social respectability. This phenomenon aligns with our model, since, alongside the stigma attached to poverty, it increases the value of being identified as middle class and intensifies competition to attain that status. At the same time, competition with the rich is attenuated, as the middle class has less to gain from emulating the rich, given its improved fallback position in terms of social status. As a result, this pattern may indirectly benefit the wealthy by allowing them to reduce conspicuous expenditures. In the same spirit, the middle class may have an incentive to raise the social status associated with poverty (that is, to rehabilitate the image of the poor) because doing so relaxes the middle class's need to invest in visible "respectability" spending to avoid downward misrecognition. These conflicting interests over the social allocation of status suggest that the dominant narrative about social classes may reflect relative power in the public sphere: groups with greater political, cultural, or media influence can more effectively shape

how poverty and affluence are framed and perceived.

The model also speaks to an apparent empirical paradox: wealthy individuals may engage in charitable giving while simultaneously supporting social arrangements and narratives that preserve class distance. In the framework developed here, these behaviors are not contradictory. In a Riley–Riley environment, transfers to the poor can reduce the rich's equilibrium need for conspicuous expenditure by intensifying status competition below. This provides a status-based rationale for why elite support may emerge for targeted charity or narrowly means-tested transfers, while broader universal or non-stigmatizing redistributive policies encounter greater resistance. More broadly, the model helps interpret social narratives that portray poverty as a marker of personal failure while celebrating middle-class respectability: such narratives do not merely reflect moral judgments, but can also reshape the equilibrium structure of status competition in ways that benefit higher-income groups.

Finally, one may ask how robust our results are to relaxing some of the model's assumptions. While two distinct locations for status competition and the middle class's simultaneous participation in both are central to the framework, the cross-location transmission mechanism itself does not depend on the restriction to three income types. Suppose that a multiplicity of income types competes for rank within each location, and that some individuals participate in both locations. An increase in the income of the poorest type steepens competition at the bottom of the status ladder in the low-status location. Because the marginal "winners" of this intensified competition also enter the high location, they bring weaker incentives to climb the top ladder, which relaxes competitive pressure for those who only compete there and generates positive spillovers at the top. The same cross-location logic implies that changing the social status attached to poverty, through stigma, can propagate beyond the bottom of the distribution via the set of individuals who frequent both locations.

Overall, our contribution sheds light on hidden mechanisms affecting preferences for income redistribution in societies characterized by class

segregation and status concerns. When status competition occurs separately in the lower and upper segments of society, strategic incentives emerge that shape support for redistributive policies and foster dynamics in which certain social groups are elevated while others are diminished.

1.9 Appendix: Proofs and additional definitions

This appendix provides the technical details and proofs of the statements presented throughout the paper.

1.9.1 Proof of Proposition 1

Proposition 1. *The game $\mathcal{G}$ admits at least one WPBE in each of the following classes: pooling–pooling, separating–separating, pooling–separating, and separating–pooling.*

Proof. We characterize and show the existence of the four equilibrium classes of this model: pooling-pooling (i.e., pooling in each location), separating-separating (i.e., separating in each location), separating-pooling (i.e., separating in location l and pooling in location h), and pooling-separating (i.e., separating in location h and pooling in location l).

Pooling–pooling. Consider a pooling–pooling candidate in which, in location ℓ, types $\mathcal{P}$ and $\mathcal{M}$ choose the same expenditure $\hat{x}_\ell$, and in location h, types $\mathcal{M}$ and $\mathcal{R}$ choose the same expenditure $\hat{x}_h$:

$$x_{\ell\mathcal{P}} = x_{\ell\mathcal{M}} = \hat{x}_\ell, \qquad x_{h\mathcal{M}} = x_{h\mathcal{R}} = \hat{x}_h.$$

On the equilibrium path, Bayes' rule implies

$$\mu_\ell(\hat{x}_\ell) = \Pr(\mathcal{M} \mid \hat{x}_\ell) = \frac{\beta}{\alpha+\beta}, \qquad \mu_h(\hat{x}_h) = \Pr(\mathcal{R} \mid \hat{x}_h) = \frac{\gamma}{\beta+\gamma},$$

so the corresponding on-path status terms are

$$s_\ell\left(\frac{\beta}{\alpha+\beta}\right) = \frac{\alpha}{\alpha+\beta}P + \frac{\beta}{\alpha+\beta}M, \qquad s_h\left(\frac{\gamma}{\beta+\gamma}\right) = \frac{\beta}{\beta+\gamma}M + \frac{\gamma}{\beta+\gamma}R.$$

Step 1a: local incentive constraints in location ℓ. Given $\hat{x}_h$, define $\overline{x}_\ell^{\,pool,\mathcal{P}}$ and $\overline{x}_\ell^{\,pool,\mathcal{M}}(\hat{x}_h)$ as the (maximal) values of $\hat{x}_\ell$ that make, respectively, $\mathcal{P}$ and $\mathcal{M}$ weakly prefer pooling at $\hat{x}_\ell$ to deviating down to 0 when off-path beliefs assign the deviator to the lower status:

$$u(I_{\mathcal{P}}) + P = u(I_{\mathcal{P}} - \overline{x}_\ell^{\,pool,\mathcal{P}}) + s_\ell\left(\frac{\beta}{\alpha+\beta}\right), \tag{1.16}$$

$$u(I_{\mathcal{M}} - \hat{x}_h) + P = u(I_{\mathcal{M}} - \hat{x}_h - \overline{x}_\ell^{\,pool,\mathcal{M}}(\hat{x}_h)) + s_\ell\left(\frac{\beta}{\alpha+\beta}\right). \tag{1.17}$$

Let

$$\overline{x}_{\ell}^{\,pool}(\hat{x}_h) \equiv \min\left\{\overline{x}_{\ell}^{\,pool,\mathcal{P}},\ \overline{x}_{\ell}^{\,pool,\mathcal{M}}(\hat{x}_h)\right\}.$$

Then for any $\hat{x}_\ell \in [0, \overline{x}_{\ell}^{\,pool}(\hat{x}_h)]$, neither $\mathcal{P}$ nor $\mathcal{M}$ has a profitable *downward* deviation in location ℓ under the beliefs specified below. (Upward deviations are never profitable since they increase expenditure without improving status.)

Step 2a: local incentive constraints in location h. Given $\hat{x}_\ell$, define $\overline{x}_{h}^{\,pool,\mathcal{M}}(\hat{x}_\ell)$ and $\overline{x}_{h}^{\,pool,\mathcal{R}}$ as the (maximal) values of $\hat{x}_h$ that make, respectively, $\mathcal{M}$ and $\mathcal{R}$ weakly prefer pooling at $\hat{x}_h$ to deviating down to 0:

$$u(I_{\mathcal{M}} - \hat{x}_\ell) + M = u(I_{\mathcal{M}} - \hat{x}_\ell - \overline{x}_{h}^{\,pool,\mathcal{M}}(\hat{x}_\ell)) + s_h\left(\frac{\gamma}{\beta+\gamma}\right), \tag{1.18}$$

$$u(I_{\mathcal{R}}) + M = u(I_{\mathcal{R}} - \overline{x}_{h}^{\,pool,\mathcal{R}}) + s_h\left(\frac{\gamma}{\beta+\gamma}\right). \tag{1.19}$$

Let

$$\overline{x}_{h}^{\,pool}(\hat{x}_\ell) \equiv \min\left\{\overline{x}_{h}^{\,pool,\mathcal{M}}(\hat{x}_\ell),\ \overline{x}_{h}^{\,pool,\mathcal{R}}\right\}.$$

Then for any $\hat{x}_h \in [0, \overline{x}_{h}^{\,pool}(\hat{x}_\ell)]$, neither $\mathcal{M}$ nor $\mathcal{R}$ has a profitable *downward* deviation in location h under the beliefs below. (Again, upward deviations are never profitable.)

Step 3a: supporting beliefs. Consider the following pessimistic off-path beliefs:

$$\mu_\ell(x_\ell) = \begin{cases} \frac{\beta}{\alpha+\beta} & \text{if } x_\ell = \hat{x}_\ell, \\ 0 & \text{if } x_\ell \neq \hat{x}_\ell, \end{cases} \qquad \mu_h(x_h) = \begin{cases} \frac{\gamma}{\beta+\gamma} & \text{if } x_h = \hat{x}_h, \\ 0 & \text{if } x_h \neq \hat{x}_h. \end{cases}$$

These beliefs are Bayes-consistent on the equilibrium path and assign any off-path observation in a location to the lower-status type in that location.

Step 4a: no cross-location profitable deviation for $\mathcal{M}$. Fix any $(\hat{x}_\ell, \hat{x}_h)$ satisfying

$$\hat{x}_\ell \in [0, \overline{x}_{\ell}^{\,pool}(\hat{x}_h)] \quad \text{and} \quad \hat{x}_h \in [0, \overline{x}_{h}^{\,pool}(\hat{x}_\ell)].$$

Any deviation $(x'_\ell, x'_h) \neq (\hat{x}_\ell, \hat{x}_h)$ by type $\mathcal{M}$ triggers $\mu_\ell(x'_\ell) = 0$ if $x'_\ell \neq \hat{x}_\ell$ and/or $\mu_h(x'_h) = 0$ if $x'_h \neq \hat{x}_h$, thus weakly lowering status in the deviated location(s). If $\mathcal{M}$ changes one component while keeping the other

fixed, the relevant bounds (1.17) and (1.18) ensure that the associated status loss cannot be offset by the consumption gain. If $\mathcal{M}$ changes both components, it loses status in both locations and the deviation is a fortiori unprofitable. Hence $\mathcal{M}$ has no profitable cross-location deviation.

Step 5a: non-emptiness. The set of pairs $(\hat{x}_\ell, \hat{x}_h)$ satisfying the above constraints is nonempty. Indeed, $(\hat{x}_\ell, \hat{x}_h) = (0, 0)$ is feasible for all types and belongs to the set since $0 \in [0, \overline{x}_\ell^{pool}(0)]$ and $0 \in [0, \overline{x}_h^{pool}(0)]$ by construction. With the supporting beliefs in Step 3, any deviation in either location is necessarily an *upward* deviation and thus strictly increases conspicuous expenditure while (weakly) lowering status, because $\mu_\ell(x_\ell) = 0$ for all $x_\ell \neq 0$ and $\mu_h(x_h) = 0$ for all $x_h \neq 0$. Hence no type has a profitable deviation from $(0, 0)$, and a pooling–pooling WPBE exists.

Therefore, for any $(\hat{x}_\ell, \hat{x}_h)$ in the nonempty set characterized above (e.g. $(0, 0)$), the strategy profile together with the beliefs defines a pooling–pooling WPBE.

Separating–separating. Consider a separating–separating candidate in which, in location ℓ, types $\mathcal{P}$ and $\mathcal{M}$ choose different expenditures and, in location h, types $\mathcal{M}$ and $\mathcal{R}$ choose different expenditures:

$$x_{\ell\mathcal{P}} = 0, \qquad x_{\ell\mathcal{M}} = \hat{x}_\ell > 0, \qquad x_{h\mathcal{M}} = 0, \qquad x_{h\mathcal{R}} = \hat{x}_h > 0.$$

Step 1b: separation in location ℓ. In a separating equilibrium, type $\mathcal{P}$ must spend 0 in location ℓ, since any positive prescribed expenditure would be strictly dominated by 0 given that status is already fixed on-path. Define the *minimal* separating expenditure for the middle type in location ℓ, $x_\ell^* > 0$, by the poor type's indifference between staying poor at 0 and mimicking the middle at x_ℓ^*:

$$u(I_\mathcal{P}) + P = u(I_\mathcal{P} - x_\ell^*) + M. \tag{1.20}$$

Define the *maximal* separating expenditure for the middle type in location ℓ, $\overline{x}_\ell > 0$, by the middle type's indifference between separating at $\overline{x}_\ell$ and deviating down to 0 (and being perceived as poor):

$$u(I_\mathcal{M}) + P = u(I_\mathcal{M} - \overline{x}_\ell) + M. \tag{1.21}$$

By strict monotonicity and continuity of $u(\cdot)$, both x_ℓ^* and $\overline{x}_\ell$ exist and satisfy $0 < x_\ell^* \leq \overline{x}_\ell$.Hence, for any $\hat{x}_\ell \in [x_\ell^*, \overline{x}_\ell]$, separation in location

ℓ can be supported under the beliefs specified below: if $\hat{x}_\ell < x^*_\ell$, type $\mathcal{P}$ would profitably mimic; if $\hat{x}_\ell > \overline{x}_\ell$, type $\mathcal{M}$ would profitably deviate down to 0.

Step 2b: separation in location h. In a separating equilibrium, type $\mathcal{M}$ must spend 0 in location h; any positive expenditure would be strictly dominated by 0 given that status is already fixed on-path. Given $\hat{x}_\ell$, define the *minimal* separating expenditure for the rich type in location h, $x^*_h(\hat{x}_\ell) > 0$, by the middle type's indifference between staying middle at 0 and mimicking the rich at $x^*_h(\hat{x}_\ell)$:

$$u(I_{\mathcal{M}} - \hat{x}_\ell) + M = u(I_{\mathcal{M}} - \hat{x}_\ell - x^*_h(\hat{x}_\ell)) + R. \tag{1.22}$$

Define the *maximal* separating expenditure for the rich type in location h, $\overline{x}_h > 0$, by the rich type's indifference between separating at $\overline{x}_h$ and deviating down to 0 (and being perceived as middle):

$$u(I_{\mathcal{R}}) + M = u(I_{\mathcal{R}} - \overline{x}_h) + R. \tag{1.23}$$

Therefore, for any $\hat{x}_h \in [x^*_h(\hat{x}_\ell), \overline{x}_h]$, separation in location h can be supported under the beliefs specified below: if $\hat{x}_h < x^*_h(\hat{x}_\ell)$, type $\mathcal{M}$ would profitably mimic; if $\hat{x}_h > \overline{x}_h$, type $\mathcal{R}$ would profitably deviate down to 0.

Step 3b: supporting beliefs. Consider the following (pessimistic) beliefs:

$$\mu_\ell(x_\ell) = \begin{cases} 1 & \text{if } x_\ell = \hat{x}_\ell, \\ 0 & \text{otherwise,} \end{cases} \qquad \mu_h(x_h) = \begin{cases} 1 & \text{if } x_h = \hat{x}_h, \\ 0 & \text{otherwise.} \end{cases}$$

These beliefs are Bayes-consistent on the equilibrium path (since $x_{\ell\mathcal{P}} = 0$ and $x_{\ell\mathcal{M}} = \hat{x}_\ell$, and $x_{h\mathcal{M}} = 0$ and $x_{h\mathcal{R}} = \hat{x}_h$), and they assign any off-path observation in a location to the lower-status type in that location.

Step 4b: no cross-location profitable deviation for $\mathcal{M}$. Under the beliefs in Step 3, any deviation by type $\mathcal{M}$ that keeps $x_\ell < \hat{x}_\ell$ yields $\mu_\ell(x_\ell) = 0$ and thus status P in location ℓ, so any expenditure $x_\ell \in (0, \hat{x}_\ell)$ is wasted relative to $x_\ell = 0$. Hence the most profitable cross-location deviation for $\mathcal{M}$ is of the form $(x'_\ell, x'_h) = (0, x'_h)$. Moreover, deviating in location h to be perceived as rich requires choosing $x'_h = \hat{x}_h$, which yields status R in h; any $x'_h \neq \hat{x}_h$ yields $\mu_h(x'_h) = 0$ and thus status M in h. Therefore, the only potentially profitable cross-location deviation is $(0, \hat{x}_h)$, and it is unprofitable if and only if

$$u(I_{\mathcal{M}} - \hat{x}_\ell) + 2M \;\geq\; u(I_{\mathcal{M}} - \hat{x}_h) + P + R. \tag{1.24}$$

Step 5b: non-emptiness. There exists at least one separating–separating WPBE. For instance, take $\hat{x}_\ell = x_\ell^*$ and $\hat{x}_h = \overline{x}_h$. By definition of $\overline{x}_h$, the rich type is indifferent between separating at $\overline{x}_h$ and deviating down to 0, while by strict concavity of $u(\cdot)$ and $I_{\mathcal{R}} > I_{\mathcal{M}}$, the middle type is strictly less willing to pay for the status gain $R - M$: letting $\overline{x}_h^{\mathcal{M}}$ solve

$$u(I_{\mathcal{M}}) + M = u(I_{\mathcal{M}} - \overline{x}_h^{\mathcal{M}}) + R,$$

we have $\overline{x}_h > \overline{x}_h^{\mathcal{M}}$, so type $\mathcal{M}$ does not find it profitable to mimic the rich even when setting $x'_\ell = 0$. Thus the strategy profile $(x_{\ell\mathcal{P}}, x_{\ell\mathcal{M}}, x_{h\mathcal{M}}, x_{h\mathcal{R}}) = (0, x_\ell^*, 0, \overline{x}_h)$ together with the beliefs in Step 3 constitutes a separating–separating WPBE.

Separating–pooling. Consider a separating–pooling candidate in which location ℓ is separating between $\mathcal{P}$ and $\mathcal{M}$, while location h is pooling between $\mathcal{M}$ and $\mathcal{R}$:

$$x_{\ell\mathcal{P}} = 0, \qquad x_{\ell\mathcal{M}} = \hat{x}_\ell > 0, \qquad x_{h\mathcal{M}} = x_{h\mathcal{R}} = \hat{x}_h \geq 0.$$

On the equilibrium path, Bayes' rule in location h implies

$$\mu_h(\hat{x}_h) = \Pr(\mathcal{R} \mid \hat{x}_h) = \frac{\gamma}{\beta + \gamma}.$$

Since $\gamma/(\beta+\gamma) \in (0,1)$ and $s_h(\cdot)$ is strictly increasing, it follows that

$$s_h(0) = M \;<\; s_h\left(\frac{\gamma}{\beta+\gamma}\right) \;<\; s_h(1) = R.$$

Step 1c: separation in location ℓ. Define $x_\ell^* > 0$ by the poor type's indifference between staying poor at 0 and mimicking the middle at x_ℓ^*:

$$u(I_{\mathcal{P}}) + P = u(I_{\mathcal{P}} - x_\ell^*) + M. \tag{1.25}$$

Given $\hat{x}_h$, define the maximal separating expenditure for $\mathcal{M}$ in location ℓ, $\overline{x}_\ell^{sep}(\hat{x}_h) > 0$, by the middle type's indifference between separating at $\overline{x}_\ell^{sep}(\hat{x}_h)$ and deviating down to 0 (and being perceived as poor):

$$u(I_{\mathcal{M}} - \hat{x}_h) + P = u(I_{\mathcal{M}} - \hat{x}_h - \overline{x}_\ell^{sep}(\hat{x}_h)) + M. \tag{1.26}$$

Hence, for any $\hat{x}_\ell \in [x_\ell^*, \overline{x}_\ell^{sep}(\hat{x}_h)]$, location ℓ can be separating under the beliefs specified below: if $\hat{x}_\ell < x_\ell^*$, type $\mathcal{P}$ would mimic; if $\hat{x}_\ell > \overline{x}_\ell^{sep}(\hat{x}_h)$, type $\mathcal{M}$ would deviate down to 0.

Step 2c: pooling in location h. Given $\hat{x}_\ell$, define $\overline{x}_h^{pool,\mathcal{M}}(\hat{x}_\ell)$ and $\overline{x}_h^{pool,\mathcal{R}}$ as the maximal values of $\hat{x}_h$ such that, respectively, $\mathcal{M}$ and $\mathcal{R}$ weakly prefer pooling at $\hat{x}_h$ to deviating down to 0 when off-path beliefs assign the deviator to the lower-status type (i.e. $\mu_h = 0$):

$$u(I_{\mathcal{M}} - \hat{x}_\ell) + s_h\left(\frac{\gamma}{\beta+\gamma}\right) = u(I_{\mathcal{M}} - \hat{x}_\ell - \overline{x}_h^{pool,\mathcal{M}}(\hat{x}_\ell)) + M, \tag{1.27}$$

$$u(I_{\mathcal{R}}) + s_h\left(\frac{\gamma}{\beta+\gamma}\right) = u(I_{\mathcal{R}} - \overline{x}_h^{pool,\mathcal{R}}) + M. \tag{1.28}$$

Let

$$\overline{x}_h^{pool}(\hat{x}_\ell) \equiv \min\left\{\overline{x}_h^{pool,\mathcal{M}}(\hat{x}_\ell),\ \overline{x}_h^{pool,\mathcal{R}}\right\}.$$

Then for any $\hat{x}_h \in [0, \overline{x}_h^{pool}(\hat{x}_\ell)]$, neither $\mathcal{M}$ nor $\mathcal{R}$ has a profitable downward deviation in location h under the beliefs specified below. (Upward deviations are never profitable since they increase expenditure without improving status.)

Step 3c: supporting beliefs. Consider the following pessimistic beliefs:

$$\mu_\ell(x_\ell) = \begin{cases} 1 & \text{if } x_\ell = \hat{x}_\ell, \\ 0 & \text{otherwise,} \end{cases} \qquad \mu_h(x_h) = \begin{cases} \frac{\gamma}{\beta+\gamma} & \text{if } x_h = \hat{x}_h, \\ 0 & \text{otherwise.} \end{cases}$$

These beliefs are Bayes-consistent on the equilibrium path and assign any off-path observation in a location to the lower-status type in that location.

Step 4c: no cross-location profitable deviation for $\mathcal{M}$. Under the beliefs in Step 3, in order to keep status M in location ℓ type $\mathcal{M}$ must choose exactly $x_\ell = \hat{x}_\ell$; any $x_\ell \neq \hat{x}_\ell$ yields $\mu_\ell(x_\ell) = 0$ and thus status P in ℓ. Similarly, in order to obtain the pooling status term $s_h(\gamma/(\beta+\gamma))$ in location h, type $\mathcal{M}$ must choose exactly $x_h = \hat{x}_h$; any $x_h \neq \hat{x}_h$ yields $\mu_h(x_h) = 0$ and thus status M in h. Therefore, any deviation $(x'_\ell, x'_h) \neq (\hat{x}_\ell, \hat{x}_h)$ necessarily lowers status in at least one location. If $\mathcal{M}$ changes only one component, the corresponding local constraints (1.26) and (1.27) ensure that the consumption gain cannot offset the induced status loss. If $\mathcal{M}$ changes both components, it loses status in both locations and the deviation is a fortiori unprofitable. Hence $\mathcal{M}$ has no profitable cross-location deviation.

Step 5c: non-emptiness. A separating–pooling WPBE exists. For instance, take $\hat{x}_h = 0$ and $\hat{x}_\ell = x_\ell^*$. Pooling in h is immediate under the

pessimistic beliefs in Step 3. Moreover, strict concavity of $u(\cdot)$ implies that if x_ℓ^* deters type $\mathcal{P}$ from mimicking, then it is strictly preferred by type $\mathcal{M}$ to deviating down to 0 (and being perceived as poor). Hence $(0, x_\ell^*, 0, 0)$ together with the beliefs in Step 3 constitutes a separating–pooling WPBE.

Pooling–separating. Consider a pooling–separating candidate in which location ℓ is pooling between $\mathcal{P}$ and $\mathcal{M}$, while location h is separating between $\mathcal{M}$ and $\mathcal{R}$:

$$x_{\ell\mathcal{P}} = x_{\ell\mathcal{M}} = \hat{x}_\ell, \qquad x_{h\mathcal{M}} = 0, \qquad x_{h\mathcal{R}} = \hat{x}_h.$$

On the equilibrium path, Bayes' rule in location ℓ implies

$$\mu_\ell(\hat{x}_\ell) = \Pr(\mathcal{M} \mid \hat{x}_\ell) = \frac{\beta}{\alpha + \beta},$$

while separation in location h implies $\mu_h(0) = 0$ and $\mu_h(\hat{x}_h) = 1$.

Step 1d: pooling in location ℓ. Given that type $\mathcal{M}$ spends $x_{h\mathcal{M}} = 0$ in location h, define $\overline{x}_\ell^{\,pool,\mathcal{P}}$ and $\overline{x}_\ell^{\,pool,\mathcal{M}}$ as the maximal values of $\hat{x}_\ell$ such that, respectively, $\mathcal{P}$ and $\mathcal{M}$ weakly prefer pooling at $\hat{x}_\ell$ to deviating down to 0 when off-path beliefs assign the deviator to the lower-status type:

$$u(I_\mathcal{P}) + P = u(I_\mathcal{P} - \overline{x}_\ell^{\,pool,\mathcal{P}}) + s_\ell\left(\frac{\beta}{\alpha + \beta}\right), \tag{1.29}$$

$$u(I_\mathcal{M}) + P = u(I_\mathcal{M} - \overline{x}_\ell^{\,pool,\mathcal{M}}) + s_\ell\left(\frac{\beta}{\alpha + \beta}\right). \tag{1.30}$$

Let

$$\overline{x}_\ell^{\,pool} \equiv \min\left\{\overline{x}_\ell^{\,pool,\mathcal{P}},\ \overline{x}_\ell^{\,pool,\mathcal{M}}\right\}.$$

Then for any $\hat{x}_\ell \in [0, \overline{x}_\ell^{\,pool}]$, neither $\mathcal{P}$ nor $\mathcal{M}$ has a profitable *downward* deviation in location ℓ under the beliefs specified below. (Upward deviations are never profitable since they increase expenditure without improving status.)

Step 2d: separation in location h. Given $\hat{x}_\ell$, define the minimal separating expenditure for the rich type in location h, $x_h^*(\hat{x}_\ell) > 0$, by the middle type's indifference between staying middle at 0 and mimicking the rich at $x_h^*(\hat{x}_\ell)$:

$$u(I_\mathcal{M} - \hat{x}_\ell) + M = u(I_\mathcal{M} - \hat{x}_\ell - x_h^*(\hat{x}_\ell)) + R. \tag{1.31}$$

Define the maximal separating expenditure for the rich type in location h, $\overline{x}_h > 0$, by the rich type's indifference between separating at $\overline{x}_h$ and deviating down to 0 (and being perceived as middle):

$$u(I_{\mathcal{R}}) + M = u(I_{\mathcal{R}} - \overline{x}_h) + R. \tag{1.32}$$

Hence, for any $\hat{x}_h \in [x_h^*(\hat{x}_\ell), \overline{x}_h]$, separation in location h can be supported under the beliefs specified below.

Step 3d: supporting beliefs. Consider the following pessimistic beliefs:

$$\mu_\ell(x_\ell) = \begin{cases} \frac{\beta}{\alpha+\beta} & \text{if } x_\ell = \hat{x}_\ell, \\ 0 & \text{otherwise,} \end{cases} \qquad \mu_h(x_h) = \begin{cases} 1 & \text{if } x_h = \hat{x}_h, \\ 0 & \text{otherwise.} \end{cases}$$

These beliefs are Bayes-consistent on the equilibrium path and assign any off-path observation in a location to the lower-status type in that location.

Step 4d: no cross-location profitable deviation for $\mathcal{M}$. Under the beliefs in Step 3, any deviation by type $\mathcal{M}$ with $x'_\ell \neq \hat{x}_\ell$ yields $\mu_\ell(x'_\ell) = 0$ and thus status P in location ℓ. Moreover, deviating in location h to be perceived as rich requires choosing $x'_h = \hat{x}_h$, while any $x'_h \neq \hat{x}_h$ yields $\mu_h(x'_h) = 0$ and thus status M in location h. Therefore, the only potentially profitable cross-location deviation for $\mathcal{M}$ is $(x'_\ell, x'_h) = (0, \hat{x}_h)$, and it is unprofitable whenever

$$u(I_{\mathcal{M}} - \hat{x}_\ell) + s_\ell\left(\frac{\beta}{\alpha+\beta}\right) + M \;\geq\; u(I_{\mathcal{M}} - \hat{x}_h) + P + R. \tag{1.33}$$

Step 5d: non-emptiness. A pooling–separating WPBE exists. For instance, take $\hat{x}_\ell = 0$ and $\hat{x}_h = \overline{x}_h$. Pooling in location ℓ at $\hat{x}_\ell = 0$ is supported by Step 1, while separation in location h is supported by Step 2. Finally, (1.33) holds for $\hat{x}_h = \overline{x}_h$ because, by strict concavity of $u(\cdot)$ and $I_{\mathcal{R}} > I_{\mathcal{M}}$, the maximum expenditure that type $\mathcal{M}$ is willing to pay to be perceived as rich in location h is strictly smaller than $\overline{x}_h$. Hence the cross-location deviation is unprofitable, and a pooling–separating WPBE exists.

□

1.9.2 Formal Definitions

Definition 4. *Riley-Riley is an outcome configuration prescribing the following strategies.*

- *The strategy of the poor is* $x_{\ell\mathcal{P}} = 0$*;*
- *The strategy of the middle is* $x_{\ell\mathcal{M}} = x_\ell^*$, $x_{h\mathcal{M}} = 0$*;*
- *The strategy of the rich is* $x_{h\mathcal{R}} = x_h^*$*;*

Where x_ℓ^* *and* x_h^* *are such that*

$$\begin{cases} u(I_{\mathcal{P}}) + P = u(I_{\mathcal{P}} - x_\ell^*) + M \\ u(I_{\mathcal{M}} - x_\ell^*) + M = u(I_{\mathcal{M}} - x_\ell^* - x_h^*) + R \end{cases}$$

Definition 5. *Riley–Non-Riley outcome is a WPBE with the following characteristics.*

- *The strategy of the poor is* $x_{\ell\mathcal{P}} = 0$*.*
- *The strategy of the middle is* $x_{\ell\mathcal{M}} = x_\ell^*$, $x_{h\mathcal{M}} = 0$*.*
- *The strategy of the rich is* $x_{h\mathcal{R}} = x_h^{**}$*.*

Where x_ℓ^* *and* x_h^{**} *are such that*

$$\begin{cases} u(I_{\mathcal{P}}) + P = u(I_{\mathcal{P}} - x_\ell^*) + M \\ u(I_{\mathcal{M}} - x_\ell^*) + 2M = u(I_{\mathcal{M}} - x_h^{**}) + P + R \end{cases}$$

1.9.3 Proof of Proposition 2

Proposition 2. *The two-location Intuitive Criterion selects a unique equilibrium outcome (up to off-path beliefs). On the equilibrium path:*

$$x_{\ell\mathcal{P}} = 0, \qquad x_{\ell\mathcal{M}} = x_\ell^*, \qquad x_{h\mathcal{M}} = 0, \qquad x_{h\mathcal{R}} = \tilde{x}_h.$$

Equivalently, the selected outcome is: (i) Riley–Riley *if* $x_h^* \geq x_h^{**}$*; and (ii)* Riley–Non-Riley *(with rich overspending in location h) if* $x_h^{**} > x_h^*$*.*

Proof. We proceed by exhaustion. We first show that any WPBE outcome involving pooling in at least one location is ruled out by the two-location Intuitive Criterion (Definition 2). We then characterize the separating–separating outcome selected by the criterion, which minimizes signaling expenditures subject to incentive compatibility and the middle type's cross-location deviation constraint.

- **Pooling–pooling.** Suppose, by contradiction, that an IC-compatible WPBE exhibits pooling in location h at some $\hat{x}_h$. Consider an off-path observation $x_h > \hat{x}_h$ in location h. By the single-crossing property established in Section 3.2, such a deviation can be weakly profitable only for the higher type in location h, i.e. $D_h(x_h) = \{\mathcal{R}\}$. Then Definition 2 requires $\mu_h(x_h) = 1$. Under this belief, the rich type strictly prefers deviating to x_h (for x_h sufficiently close to $\hat{x}_h$), contradicting pooling in h. Therefore, pooling in h (and hence pooling–pooling) cannot survive the Intuitive Criterion.

- **Separating–pooling.** This class features pooling in location h at some $\hat{x}_h$. The same argument as above applies: for any off-path $x_h > \hat{x}_h$, single-crossing implies $D_h(x_h) = \{\mathcal{R}\}$, hence the Intuitive Criterion requires $\mu_h(x_h) = 1$, which makes the deviation profitable for $\mathcal{R}$. Thus separating–pooling outcomes are ruled out.

- **Pooling–separating.** Suppose, by contradiction, that an IC-compatible WPBE exhibits pooling in location ℓ at some $\hat{x}_\ell$. Consider an off-path observation $x'_\ell \in (I_\mathcal{P}, I_\mathcal{M}]$ in location ℓ (and $x'_\ell \neq \hat{x}_\ell$). Such a deviation is infeasible for type $\mathcal{P}$, while it is feasible for type $\mathcal{M}$ (by setting $x_h = 0$ if needed). Hence $D_\ell(x'_\ell) = \{\mathcal{M}\}$ in the sense of Definition 2, so the two-location Intuitive Criterion requires $\mu_\ell(x'_\ell) = 1$. Under this belief, type $\mathcal{M}$ can profitably deviate to some x'_ℓ (arbitrarily close to $\hat{x}_\ell$ whenever $\hat{x}_\ell < I_\mathcal{P}$), contradicting pooling in ℓ. Therefore, pooling–separating outcomes are ruled out.

- **Separating–separating.** We are left with outcomes that are separating in both locations. In location ℓ, the Intuitive Criterion selects the *minimal* separating expenditure (Riley outcome): any separating profile with $x_{\ell\mathcal{M}} > x^*_\ell$ admits a slightly smaller deviation $x'_\ell \in (x^*_\ell, x_{\ell\mathcal{M}})$ that cannot be rationalized for type $\mathcal{P}$ but can be profitable for type $\mathcal{M}$; hence $D_\ell(x'_\ell) = \{\mathcal{M}\}$ and the Intuitive Criterion requires $\mu_\ell(x'_\ell) = 1$, making the deviation profitable and contradicting $x_{\ell\mathcal{M}} > x^*_\ell$. Thus $x_{\ell\mathcal{P}} = 0$ and $x_{\ell\mathcal{M}} = x^*_\ell$.

  In location h, separation requires the rich type to choose an expenditure that deters imitation by the middle type. If the standard Riley level x^*_h already blocks the middle type's best cross-location deviation (i.e. if $x^*_h \geq x^{**}_h$), then the same minimality argument implies $x_{h\mathcal{R}} = x^*_h$. If instead $x^{**}_h > x^*_h$, then any $x_{h\mathcal{R}} < x^{**}_h$ makes the cross-location deviation by $\mathcal{M}$ profitable, while any $x_{h\mathcal{R}} > x^{**}_h$ can

be reduced without violating incentive compatibility and is eliminated by the Intuitive Criterion's minimality logic. Hence $x_{h\mathcal{R}} = x_h^{**}$ in this case.

Therefore, the Intuitive Criterion selects $x_{\ell\mathcal{P}} = 0$, $x_{\ell\mathcal{M}} = x_\ell^*$, $x_{h\mathcal{M}} = 0$, and $x_{h\mathcal{R}} = \tilde{x}_h \equiv \max\{x_h^*, x_h^{**}\}$, which proves Proposition 2.

□

1.9.4 Proof of Lemma 1

Lemma 1. *Let $\tau \in (0,1)$, let $x_{\mathcal{R}}^{one}$ solve* (1.6)*, and let $\tilde{x}_h(\tau)$ solve* (1.9)*. Then*

$$\tilde{x}_h(\tau) = \max\{x_h^*(\tau), x_h^{**}(\tau)\} < x_{\mathcal{R}}^{one}.$$

Proof. We prove the claim by comparing the two segregation thresholds $x_h^*(\tau)$ and $x_h^{**}(\tau)$ with the one-location rich expenditure $x_{\mathcal{R}}^{one}$.

We first note that $x_\ell^* = x_{\mathcal{M}}^{one}$. Indeed, (1.1) and (1.5) pin down the *same* indifference condition for the middle class: in location ℓ the middle class chooses the smallest expenditure that makes the poor indifferent between remaining poor and mimicking the middle. Since u is strictly increasing, this cutoff is uniquely defined, so the two equations imply the same solution.

Step 1: $x_h^*(\tau) < x_{\mathcal{R}}^{one}$. The benchmark condition (1.6) characterizes $x_{\mathcal{R}}^{one}$ as the expenditure that just prevents the middle class from profitably mimicking the rich when *all* status competition takes place in one location. Using $x_\ell^* = x_{\mathcal{M}}^{one}$, we can rewrite (1.6) as

$$u(I_{\mathcal{M}} - x_\ell^*) \;-\; u\big(I_{\mathcal{M}} - x_\ell^* - \big(x_{\mathcal{R}}^{one} - x_\ell^*\big)\big) = R - M.$$

This expression makes the comparison transparent: the left-hand side is exactly the utility loss from increasing conspicuous spending from x_ℓ^* to $x_{\mathcal{R}}^{one}$, and the right-hand side is the status gain from moving from M to R.

Under segregation, the deviation the rich want to deter is the middle class mimicking them *in location h only*. Condition (1.7) reads

$$u(I_{\mathcal{M}} - x_\ell^*) \;-\; u(I_{\mathcal{M}} - x_\ell^* - x_h^*(\tau)) = (1-\tau)(R - M).$$

The key difference is the factor $(1-\tau)$: because the middle class spends only a fraction $(1-\tau)$ of its time in h, it enjoys the higher status R in h only

for that fraction of time. Hence, relative to the benchmark, the *effective* status gain from mimicking in h falls from $(R-M)$ to $(1-\tau)(R-M)$. Now, since $\tau \in (0,1)$ implies $(1-\tau)(R-M) < (R-M)$, compare the two defining equalities for $x_{\mathcal{R}}^{one}$ and $x_h^*(\tau)$.

In both equations, the left-hand side has the form

$$u(I_{\mathcal{M}} - x_\ell^*) - u(I_{\mathcal{M}} - x_\ell^* - z),$$

for some expenditure increment z. The first term is the same in the two equations. Moreover, because u is strictly increasing, the expression $u(I_{\mathcal{M}} - x_\ell^* - z)$ is *strictly decreasing* in z (higher z leaves less consumption). Therefore the difference $u(I_{\mathcal{M}} - x_\ell^*) - u(I_{\mathcal{M}} - x_\ell^* - z)$ is *strictly increasing* in z.

In the benchmark, the difference equals $R - M$ when $z = x_{\mathcal{R}}^{one} - x_\ell^*$. Under segregation, the same type of difference must equal the smaller quantity $(1-\tau)(R-M)$ when $z = x_h^*(\tau)$. Since the left-hand side increases with z, obtaining a smaller value requires a smaller increment. Hence

$$x_h^*(\tau) < x_{\mathcal{R}}^{one} - x_\ell^*.$$

Finally, since $x_\ell^* > 0$, we have $x_{\mathcal{R}}^{one} - x_\ell^* < x_{\mathcal{R}}^{one}$, so

$$x_h^*(\tau) < x_{\mathcal{R}}^{one}.$$

Intuitively, because the middle class spends only a fraction $(1-\tau)$ of its time in h, the status gain from mimicking the rich in h is smaller. As a result, the rich can deter imitation with a lower conspicuous signal.

Step 2: $x_h^{**}(\tau) < x_{\mathcal{R}}^{one}$. We now turn to the second candidate level $x_h^{**}(\tau)$, which comes from the constraint that prevents profitable deviations by the middle class when it can split its status competition across locations.

Combining (1.6) with (1.8) yields

$$u(I_{\mathcal{M}} - x_h^{**}(\tau)) = u(I_{\mathcal{M}} - x_{\mathcal{R}}^{one}) + \tau\,(R-P).$$

This equation already reveals the direction of the comparison. Since $R > P$ and $\tau \in (0,1)$, we have $\tau(R-P) > 0$, so the right-hand side is strictly larger than $u(I_{\mathcal{M}} - x_{\mathcal{R}}^{one})$. Hence

$$u(I_{\mathcal{M}} - x_h^{**}(\tau)) > u(I_{\mathcal{M}} - x_{\mathcal{R}}^{one}).$$

Because u is strictly increasing, the inequality must hold also at the level of consumption:

$$I_{\mathcal{M}} - x_h^{**}(\tau) > I_{\mathcal{M}} - x_{\mathcal{R}}^{one},$$

which implies

$$x_h^{**}(\tau) < x_{\mathcal{R}}^{one}.$$

Intuitively: this constraint is also weaker than the benchmark one-location constraint, because the middle class now also derives some status utility from ℓ (for a fraction τ of time), which reduces the pressure the rich face in h and lowers the expenditure needed to maintain separation.

Conclusion. We have shown that both $x_h^*(\tau)$ and $x_h^{**}(\tau)$ lie strictly below $x_{\mathcal{R}}^{one}$. Therefore their maximum also lies strictly below $x_{\mathcal{R}}^{one}$:

$$\tilde{x}_h(\tau) = \max\{x_h^*(\tau), x_h^{**}(\tau)\} < x_{\mathcal{R}}^{one}.$$

□

1.9.5 Proof of Proposition 3

Proposition 3. *A marginal increase in the poor's income affects the rich's equilibrium conspicuous spending in location h, depending on the equilibrium configuration:*

(i) Riley–Riley. *A marginal increase in the income of the poor reduces the rich's equilibrium conspicuous spending in h, so the rich are better off.*

(ii) Riley–Non-Riley *A marginal increase in the income of the poor increases the rich's equilibrium conspicuous spending in h, so the rich are worse off.*

Proof. Let the poor's income increase from $I_{\mathcal{P}}$ to $I_{\mathcal{P}} + \delta$, with $\delta > 0$ arbitrarily small. Assume the equilibrium configuration is locally unchanged for δ small enough (e.g. Riley–Riley if $x_h^* > x_h^{**}$ at $\delta = 0$, and Riley–Non-Riley if $x_h^{**} > x_h^*$ at $\delta = 0$).[5] We show how the objects x_ℓ^*, x_h^*, and x_h^{**} move with $I_{\mathcal{P}}$.

Step 1: x_ℓ^* increases in $I_{\mathcal{P}}$. Define $x_\ell^* = x_\ell^*(I_{\mathcal{P}})$ by the poor's indifference condition:

$$u(I_{\mathcal{P}} - x_\ell^*(I_{\mathcal{P}})) + M \;=\; u(I_{\mathcal{P}}) + P. \tag{1.34}$$

Let

$$\Phi(x_\ell, I_{\mathcal{P}}) := u(I_{\mathcal{P}} - x_\ell) - u(I_{\mathcal{P}}) + (M - P).$$

[5] We focus on the generic case $x_h^* \neq x_h^{**}$, which ensures the equilibrium configuration is locally well-defined. If instead $x_h^* = x_h^{**}$, then a marginal increase in $I_{\mathcal{P}}$ implies x_h^{**} becomes binding because x_h^* weakly decreases while x_h^{**} strictly increases (Steps 2–3), so $\tilde{x}_h$ increases as in the Riley–Non-Riley case.

Then $\Phi(x_\ell^*(I_\mathcal{P}), I_\mathcal{P}) = 0$. By strict monotonicity and concavity of u,

$$\frac{\partial \Phi}{\partial x_\ell}(x_\ell, I_\mathcal{P}) = -u'(I_\mathcal{P} - x_\ell) < 0, \qquad \frac{\partial \Phi}{\partial I_\mathcal{P}}(x_\ell, I_\mathcal{P}) = u'(I_\mathcal{P} - x_\ell) - u'(I_\mathcal{P}) \geq 0.$$

Hence, by the implicit function theorem,

$$\frac{\partial x_\ell^*}{\partial I_\mathcal{P}} = -\frac{\Phi_{I_\mathcal{P}}}{\Phi_{x_\ell}} \geq 0, \tag{1.35}$$

so a marginal increase in $I_\mathcal{P}$ weakly increases x_ℓ^*.

Step 2: x_h^* decreases in x_ℓ^*, hence in $I_\mathcal{P}$. Define $x_h^* = x_h^*(x_\ell)$ by the middle class indifference condition:

$$u(I_\mathcal{M} - x_\ell - x_h^*(x_\ell)) + R \;=\; u(I_\mathcal{M} - x_\ell) + M. \tag{1.36}$$

Let

$$\Psi(x_h, x_\ell) := u(I_\mathcal{M} - x_\ell - x_h) - u(I_\mathcal{M} - x_\ell) + (R - M).$$

Then $\Psi(x_h^*(x_\ell), x_\ell) = 0$. Again by strict monotonicity and concavity,

$$\frac{\partial \Psi}{\partial x_h}(x_h, x_\ell) = -u'(I_\mathcal{M} - x_\ell - x_h) < 0,$$

and

$$\frac{\partial \Psi}{\partial x_\ell}(x_h, x_\ell) = -u'(I_\mathcal{M} - x_\ell - x_h) + u'(I_\mathcal{M} - x_\ell) \leq 0.$$

Therefore, by the implicit function theorem,

$$\frac{\partial x_h^*}{\partial x_\ell} = -\frac{\Psi_{x_\ell}}{\Psi_{x_h}} \leq 0. \tag{1.37}$$

Combining (1.35) and (1.37) implies that a marginal increase in $I_\mathcal{P}$ weakly decreases x_h^*.

Step 3: x_h^{**} increases in x_ℓ^*, hence in $I_\mathcal{P}$. Define $x_h^{**} = x_h^{**}(x_\ell)$ by the rich's indifference condition that pins down overspending:

$$u(I_\mathcal{M} - x_h^{**}(x_\ell)) + P + R \;=\; u(I_\mathcal{M} - x_\ell) + 2M. \tag{1.38}$$

Let

$$\Omega(x_h, x_\ell) := u(I_\mathcal{M} - x_h) - u(I_\mathcal{M} - x_\ell) + (P + R - 2M).$$

Then $\Omega(x_h^{**}(x_\ell), x_\ell) = 0$. We have

$$\frac{\partial \Omega}{\partial x_h}(x_h, x_\ell) = -u'(I_{\mathcal{M}} - x_h) < 0, \qquad \frac{\partial \Omega}{\partial x_\ell}(x_h, x_\ell) = +u'(I_{\mathcal{M}} - x_\ell) > 0,$$

so the implicit function theorem yields

$$\frac{\partial x_h^{**}}{\partial x_\ell} = -\frac{\Omega_{x_\ell}}{\Omega_{x_h}} > 0. \tag{1.39}$$

Combining (1.35) and (1.39) implies that a marginal increase in $I_{\mathcal{P}}$ strictly increases x_h^{**}.

Step 4: conclude by equilibrium configuration. Recall $\tilde{x}_h = \max\{x_h^*, x_h^{**}\}$.

(i) If the equilibrium is Riley–Riley (locally), then $\tilde{x}_h = x_h^*$ and Step 2 implies $\tilde{x}_h$ decreases in $I_{\mathcal{P}}$; the rich spend less in h and are better off.

(ii) If the equilibrium is Riley–Non-Riley (locally), then $\tilde{x}_h = x_h^{**}$ and Step 3 implies $\tilde{x}_h$ increases in $I_{\mathcal{P}}$; the rich spend more in h and are worse off.

□

1.9.6 Proof of Corollary 1

Corollary 1. *In the Riley–Riley case, the rich support a positive transfer when this status-based gain offsets the direct monetary loss, namely*

$$\frac{\gamma}{\alpha} \geq \frac{1}{\left(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x_\ell^*)}\right)\left(1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}\right)}.$$

Proof. We interpret the transfer as a *voluntary donation* by the rich. Each rich individual gives a transfer of size $t \geq 0$ to a charity, which redistributes the total amount so that each poor individual receives the lump-sum transfer

$$\Delta I_{\mathcal{P}}(t) = \frac{\gamma}{\alpha}\, t.$$

We study whether the rich would like to increase t starting from $t = 0$.

In the Riley–Riley configuration, the rich's payoff is

$$U_{\mathcal{R}}(t) = u\big(I_{\mathcal{R}} - t - x_h(t)\big) + R,$$

where $x_h(t)$ denotes the equilibrium conspicuous expenditure in location h induced by the donation. The rich support a positive donation (locally) if a marginal increase in t raises their payoff, that is, if

$$U'_{\mathcal{R}}(0) \geq 0.$$

Differentiating yields

$$U'_{\mathcal{R}}(0) = u'\big(I_{\mathcal{R}} - x_h^*\big)\Big(-1 - x'_h(0)\Big).$$

Since $u' > 0$, this condition is equivalent to

$$-x'_h(0) \geq 1. \tag{1.40}$$

Condition (1.40) has a natural interpretation: a marginal donation of one unit of income is privately desirable for the rich if and only if it reduces their equilibrium wasteful conspicuous spending by at least one unit.

To compute $x'_h(0)$, note that a donation affects x_h only through its effect on the middle class' equilibrium spending in ℓ:

$$t \;\Rightarrow\; \Delta I_{\mathcal{P}}(t) \;\Rightarrow\; x_\ell(t) \;\Rightarrow\; x_h(t).$$

Hence, by the chain rule,

$$x'_h(0) = \frac{dx_h}{dx_\ell}\Big|_{(x_\ell^*, x_h^*)} \cdot \frac{dx_\ell}{d\Delta I_{\mathcal{P}}}\Big|_{\Delta I_{\mathcal{P}}=0} \cdot \frac{d\Delta I_{\mathcal{P}}}{dt}\Big|_{t=0}.$$

Step 1: response of x_ℓ to the transfer. The equilibrium cutoff in ℓ is defined by

$$u(I_{\mathcal{P}} + \Delta I_{\mathcal{P}}) + P = u(I_{\mathcal{P}} + \Delta I_{\mathcal{P}} - x_\ell) + M.$$

Differentiating at $\Delta I_{\mathcal{P}} = 0$ (so $x_\ell = x_\ell^*$) yields

$$u'(I_{\mathcal{P}}) = u'(I_{\mathcal{P}} - x_\ell^*)\Big(1 - \frac{dx_\ell}{d\Delta I_{\mathcal{P}}}\Big),$$

and therefore

$$\frac{dx_\ell}{d\Delta I_{\mathcal{P}}}\Big|_0 = 1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x_\ell^*)} \;>\; 0. \tag{1.41}$$

Step 2: response of x_h to x_ℓ in the Riley–Riley case. Here the rich's equilibrium spending in h satisfies

$$u(I_{\mathcal{M}} - x_\ell) + M = u(I_{\mathcal{M}} - x_\ell - x_h) + R.$$

Differentiating at $(x_\ell, x_h) = (x_\ell^*, x_h^*)$ gives

$$-u'(I_{\mathcal{M}} - x_\ell^*) = -u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)\Big(1 + \frac{dx_h}{dx_\ell}\Big),$$

so

$$\frac{dx_h}{dx_\ell}\Big|_{(x_\ell^*, x_h^*)} = \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)} - 1 \;<\; 0. \tag{1.42}$$

Step 3: mechanical effect of t on the transfer. Since $\Delta I_{\mathcal{P}}(t) = \frac{\gamma}{\alpha} t$,

$$\frac{d\Delta I_{\mathcal{P}}}{dt}\Big|_0 = \frac{\gamma}{\alpha}. \tag{1.43}$$

Combining (1.41)–(1.43) yields

$$x_h'(0) = \Big(\frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)} - 1\Big)\Big(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x_\ell^*)}\Big) \cdot \frac{\gamma}{\alpha}.$$

Substituting this expression into (1.40) and dividing by $I_{\mathcal{R}} > 0$ gives

$$\frac{\gamma}{\alpha}\Big(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x_\ell^*)}\Big)\Big(1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}\Big) \;\geq\; 1,$$

which is equivalent to (1.10). This proves the corollary. □

1.9.7 Proof of Corollary 2

Corollary 2. *In the Riley–Riley case, fix $\tau \in (0,1)$. Consider a balanced-budget policy indexed by $\kappa > 0$ that taxes each middle-class and rich individual a lump-sum amount κ and rebates the proceeds equally to the poor. Thus, each poor individual receives the subsidy*

$$\Delta I_{\mathcal{P}}(\kappa) = \frac{1-\alpha}{\alpha}\,\kappa.$$

Then there exists $\bar{\kappa} > 0$ such that a policy with any $\kappa \in (0, \bar{\kappa}]$ is supported by the majority of the population if

$$\left(1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}\right)\left(1 + \frac{1-\alpha}{\alpha}\Big(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x_\ell^*)}\Big)\right) \;\geq\; 1.$$

$$\textit{and} \qquad \alpha + \gamma \geq \frac{1}{2}.$$

Proof. We compare each class' payoff under the policy and then apply majority voting. Consider throughout the proof that $\beta + \gamma = 1 - \alpha$ by definition.

Step 1: Poor support. The policy increases the poor's consumption by $\Delta I_{\mathcal{P}}(\kappa) = \frac{\beta+\gamma}{\alpha}\kappa$. Since the poor still spend zero in equilibrium, their payoff is

$$U_{\mathcal{P}}(\kappa) = u\left(I_{\mathcal{P}} + \frac{\beta+\gamma}{\alpha}\kappa\right) + P,$$

which strictly increases in κ because $u' > 0$. Hence the poor vote in favor.

Step 2: Middle class oppose. The middle class pays κ and faces tougher competition in location ℓ because the poor become richer. Therefore, the middle class must increase its separating expenditure in ℓ from x_ℓ^* to some $x_\ell(\kappa) > x_\ell^*$ (by the same logic as in Proposition 3). Since the policy reduces middle-class consumption (directly through κ and indirectly through higher x_ℓ) while leaving their status payoff at M, the middle class strictly lose and vote against any $\kappa > 0$.

Step 3: Rich support depends on the equilibrium configuration. Write the rich's payoff as

$$U_{\mathcal{R}}(\kappa) = u\big(I_{\mathcal{R}} - \kappa - \tilde{x}_h(\kappa)\big) + R,$$

where $\tilde{x}_h(\kappa)$ is the equilibrium conspicuous spending in h after the policy. The rich support the policy locally if $U'_{\mathcal{R}}(0) \geq 0$, i.e.,

$$-\tilde{x}'_h(0) \geq 1. \tag{1.44}$$

Riley–Riley. In this case $\tilde{x}_h(\kappa) = x_h(\kappa)$ solves

$$u(I_{\mathcal{M}} - \kappa - x_\ell(\kappa)) + M = u(I_{\mathcal{M}} - \kappa - x_\ell(\kappa) - x_h(\kappa)) + R.$$

Differentiating at $\kappa = 0$ yields

$$x'_h(0) = \Big(\frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)} - 1\Big)\Big(1 + x'_\ell(0)\Big).$$

Next, $x_\ell(\kappa)$ is pinned down by the poor's indifference condition in ℓ,

$$u(I_{\mathcal{P}} + \Delta I_{\mathcal{P}}(\kappa)) + P = u(I_{\mathcal{P}} + \Delta I_{\mathcal{P}}(\kappa) - x_\ell(\kappa)) + M,$$

with $\Delta I_{\mathcal{P}}(\kappa) = \frac{\beta+\gamma}{\alpha}\kappa$. Differentiating at $\kappa = 0$ gives

$$x'_\ell(0) = \frac{\beta+\gamma}{\alpha}\Big(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x^*_\ell)}\Big).$$

Substituting these expressions into (1.44) and rearranging delivers condition (1.11).

Riley–Non-Riley. In this case, Proposition 3 implies that making the poor richer raises the rich's required signal in h (to deter cross-location deviations), and the tax κ also reduces their consumption directly. Thus $U'_{\mathcal{R}}(0) < 0$ and the rich oppose the policy locally.

Step 4: Majority. Steps 1–3 imply that the poor and the rich vote in favor, and the middle class votes against. Therefore, the policy wins a majority if the poor can form a majority coalition with the rich, which requires $\alpha + \gamma \geq \frac{1}{2}$. □

1.9.8 Proof of Proposition 4

Proposition 4. *Introducing poverty stigma ($B > 0$) reduces the equilibrium conspicuous spending of the rich in both the Riley–Riley and the Riley–Non-Riley configurations.*

Proof. With stigma, the poor's indifference condition in location ℓ becomes

$$u(I_{\mathcal{P}} + \Delta I_{\mathcal{P}}) + (P - B) = u(I_{\mathcal{P}} + \Delta I_{\mathcal{P}} - x_\ell) + M,$$

where $\Delta I_{\mathcal{P}} = \frac{\gamma}{\alpha} t I_{\mathcal{R}}$ is the income transfer. Relative to a non-stigmatizing policy, the left-hand side is reduced by B, so the middle class must increase its separating expenditure in ℓ. The resulting change can be written as

$$x^{*'}_\ell - x^*_\ell = \frac{\gamma}{\alpha} t I_{\mathcal{R}} \left(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x^*_\ell)}\right) + \frac{B}{u'(I_{\mathcal{P}} - x^*_\ell)} \;\geq\; 0.$$

In a Riley–Riley configuration, a higher x_ℓ raises the opportunity cost for the middle class of mimicking the rich in location h. As a result, the rich can sustain separation with lower conspicuous spending, and

$$x^{*'}_h - x^*_h = (x^{*'}_\ell - x^*_\ell)\left(\frac{u'(I_{\mathcal{M}} - x^*_\ell)}{u'(I_{\mathcal{M}} - x^*_\ell - x^*_h)} - 1\right) \;\leq\; 0.$$

Stigma therefore amplifies the reduction in the rich's wasteful spending.

In a Riley–Non-Riley configuration, stigma affects the rich through two channels: it intensifies competition in ℓ, but it also makes cross-location deviations by the middle class less attractive. Differentiating the relevant indifference conditions yields

$$\frac{dx_h^{**}}{dP} < 0,$$

so stigma again reduces the conspicuous spending required of the rich. Hence, in both configurations, poverty stigma generates a positive status-based effect for the rich. □

1.9.9 Proof of Corollary 3

Corollary 3. Consider a redistributive policy that transfers a fraction t of the rich's income to the poor and induces poverty stigma of magnitude $B > 0$.

(i) The poor accept the policy if and only if the monetary gain outweighs the stigma cost,

$$B < u\Big(I_{\mathcal{P}} + \frac{\gamma}{\alpha} t I_{\mathcal{R}}\Big) - u(I_{\mathcal{P}}).$$

(ii) In a Riley–Riley configuration, the rich support the policy if and only if stigma is sufficiently strong to offset the direct monetary cost of the transfer,

$$B > tI_{\mathcal{R}} \left[\frac{u'(I_{\mathcal{P}} - x_\ell^*)}{1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}} - \frac{\gamma}{\alpha}\big(u'(I_{\mathcal{P}} - x_\ell^*) - u'(I_{\mathcal{P}})\big) \right].$$

A stigmatizing transfer is jointly supported by the rich and the poor if and only if the interval defined by (1.13) and (1.14) is non-empty, which requires

$$\frac{\gamma}{\alpha} > \frac{1}{1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}}.$$

Proof. We derive the acceptance conditions of the poor and the rich separately and then characterize when they overlap.

Poor. Under a stigmatizing redistributive policy, each poor individual receives the monetary transfer $\Delta I_{\mathcal{P}} = \frac{\gamma}{\alpha} t I_{\mathcal{R}}$ but suffers a status loss B. Their payoff after the policy is therefore

$$U_{\mathcal{P}}^{\text{post}} = u\Big(I_{\mathcal{P}} + \frac{\gamma}{\alpha} t I_{\mathcal{R}}\Big) + (P - B),$$

while their pre-policy payoff is $u(I_{\mathcal{P}}) + P$. The poor accept the policy if and only if $U_{\mathcal{P}}^{\text{post}} \geq u(I_{\mathcal{P}}) + P$, which is equivalent to

$$B < u\Big(I_{\mathcal{P}} + \frac{\gamma}{\alpha} t I_{\mathcal{R}}\Big) - u(I_{\mathcal{P}}),$$

yielding condition (1.13).

Rich (Riley–Riley configuration). In the Riley–Riley case, the rich's payoff after the policy is

$$U_{\mathcal{R}}^{\text{post}} = u\big(I_{\mathcal{R}} - t I_{\mathcal{R}} - x_h^{*'}\big) + R,$$

where $x_h^{*'}$ denotes the equilibrium conspicuous expenditure in h after redistribution and stigma. Before the policy, the rich obtain $u(I_{\mathcal{R}} - x_h^*) + R$. Hence, the rich support the policy if and only if

$$u\big(I_{\mathcal{R}} - t I_{\mathcal{R}} - x_h^{*'}\big) \geq u(I_{\mathcal{R}} - x_h^*).$$

Using a first-order approximation around $t = 0$ and $B = 0$, this condition can be written as

$$-t I_{\mathcal{R}} - (x_h^{*'} - x_h^*) \;\geq\; 0.$$

From Proposition 4, the change in the rich's conspicuous spending induced by redistribution and stigma is

$$x_h^{*'} - x_h^* = \left(\frac{\gamma}{\alpha} t I_{\mathcal{R}} \left(1 - \frac{u'(I_{\mathcal{P}})}{u'(I_{\mathcal{P}} - x_\ell^*)}\right) + \frac{B}{u'(I_{\mathcal{P}} - x_\ell^*)}\right)\left(\frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)} - 1\right),$$

which is weakly negative. Substituting this expression into the inequality above and rearranging yields condition (1.14).

Joint support. A stigmatizing redistributive policy is jointly supported by the poor and the rich if and only if there exists a value of B satisfying

both (1.13) and (1.14). This requires the upper bound in (1.13) to exceed the lower bound in (1.14), which is equivalent to

$$\frac{\gamma}{\alpha} > \frac{1}{1 - \frac{u'(I_{\mathcal{M}} - x_\ell^*)}{u'(I_{\mathcal{M}} - x_\ell^* - x_h^*)}}.$$

This establishes condition (1.15). □

1.10 Appendix: Social status utility and outcome configurations

We illustrate how social status utility affects the configuration of the unique equilibrium that survives the Intuitive Criterion.

The crucial condition determining which outcome configuration arises is whether it is profitable to present oneself as affluent in the high location, even if this entails being recognized as poor in the low location. If this is the case, the Riley–Non-Riley outcome obtains, since the middle class can profitably deviate from a Riley–Riley outcome. This occurs when R is relatively large or P is relatively small, leading to the failure of condition (1.2).

We present an example showing how, *ceteris paribus*, the social status utility associated with being recognized as rich determines the configuration of the unique equilibrium.

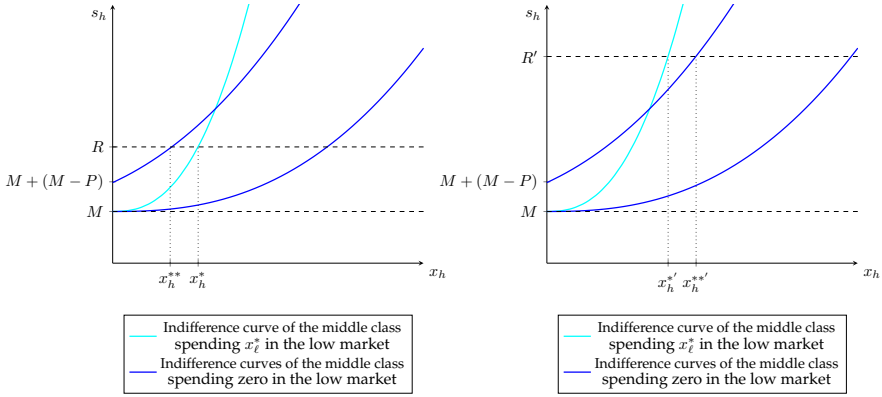


Figure 4: *Left panel: Location h in the Riley–Riley equilibrium. Right panel: Location h in the Riley–Non-Riley equilibrium.*

The left panel of Figure 4 illustrates the indifference curves of the middle class, which allow us to visualize the equilibrium level of expenditure in the high market. The lightly colored curve represents the indifference curve of the middle class when spending x_ℓ^* in the low location,

as in the Riley–Riley outcome. The darker curves are flatter because they depict the indifference curves of the middle class when spending zero in the low market.

The upper darker curve identifies the combinations (s_h, x_h) that make the middle class indifferent between (i) spending x_ℓ^* in the low market and obtaining status utility $2M$, and (ii) spending x_h^{**} in the high market and obtaining status utility $P + R$. The equilibrium choice of the rich is the larger of x_h^* and x_h^{**}, since the smaller value would not prevent profitable imitation by the middle class. In the left panel of Figure 4, we observe a Riley–Riley outcome in the high market, as $x_h^{**} < x_h^*$.

Consider now a case in which the status utility of being recognized as rich increases to $R' > R$, *ceteris paribus*. If R' is sufficiently large, then, given the shape of the indifference curves, x_h^{**} shifts to the right and exceeds x_h^*. As a result, x_h^{**} becomes the equilibrium expenditure of the rich in the high market, as illustrated in the right panel of Figure 4.

Chapter 2

A Belief-based Case for Competition in Costly Talk

This chapter is based on joint work with Federico Vaccari. Minor editorial assistance, limited to text refinement, was performed using AI-based tools.

2.1 Introduction

It is widely believed that competition among informed senders improves the quality of information available to otherwise uninformed decision makers (Gentzkow and Shapiro, 2008).

This intuition is easily proved in classic informational frameworks. In cheap talk games, if adding more senders lowers the receiver's expected utility, the receiver can simply ignore them. These extra senders then babble without affecting outcomes, and the receiver's welfare returns to its original level. On the other hand, in games with verifiable information, competition leads to full revelation (Milgrom and Roberts, 1986). Hence, in those settings, competition cannot make the receiver worse off than monopoly of information.

However, this logic fails once communication becomes costly. When misreporting entails a cost, senders cannot simply be ignored, as doing so would induce the ignored senders to tell the truth, altering equilib-

rium incentives. While Ottaviani and Squintani (2006) show that equilibria leading to full revelation exist with only one sender, little is known about the role of competition in this intermediate framework. For the first time, Vaccari (2023a) characterize equilibria in costly talk under competition between senders, proving the existence of pooling strategies when the state space is continuous and the decision set is binary. This highlights the importance of understanding the role of competition also in this context, since costly talk introduces a continuum between cheap talk and verifiable disclosure and generates a multiplicity of equilibria. In such an environment, some monopoly equilibria can be more informative than those arising under competition, leaving open the question of whether competition truly benefits the receiver.

This paper provides an answer to that question.

A useful way to interpret the mechanism is to think of competing experts advising a decision maker. Consider, for example, two policy advisors recommending whether to adopt a costly intervention. Each advisor observes relevant information about the underlying state, but may prefer a different course of action and therefore has an incentive to exaggerate the case for her preferred policy. At the same time, exaggeration is not free: a report that is too far from the truth may damage the advisor's credibility and make the recommendation less persuasive. When only one advisor is present, strategic distortion can persist because the receiver has no competing message against which to evaluate the report. Under competition, instead, each advisor knows that excessive exaggeration is more likely to be exposed by the other informed sender. This strategic interaction disciplines misreporting and improves the quality of information available to the decision maker.

This intuition guides the analysis in the paper. I show that, under a novel behavioral refinement, the *Laplacian Criterion*, competition improves the receiver's welfare relative to monopoly. The refinement is grounded in the Principle of Insufficient Reason, or Principle of Indifference (Laplace, 1812), which posits that when an agent faces uncertainty without a clear probabilistic prior, she assigns equal probability to all possible outcomes. I formalize this behavioral rule as a belief refinement

and use it to reduce the multiplicity of equilibria in costly talk games.

Importantly, the behavioral appeal of the Principle of Insufficient Reason is particularly strong in this environment. In costly talk games, the receiver often faces off-path messages for which the equilibrium provides little guidance on how beliefs should be formed. Standard refinements in signaling games typically restrict beliefs by ruling out implausible deviations based on payoff dominance. However, in this setting, the presence of multiple senders and costly misreporting generates situations in which several deviations remain observationally equivalent and cannot be ranked using standard dominance arguments. In these cases, assigning equal probability across feasible underlying states provides a natural and disciplined way to complete the model. The Laplacian Criterion can thus be interpreted as a complementary refinement: rather than eliminating equilibria through stronger rationality assumptions, it selects among them by imposing a minimal and behaviorally grounded structure on beliefs when strategic uncertainty is irreducible.

My main contribution is to establish that competition between informed senders benefits an uninformed receiver even when communication is costly. This result bridges the theoretical gap between the literatures on cheap talk and verifiable disclosure, showing that the welfare advantages of competition persist in intermediate environments. Methodologically, I introduce a behavioral equilibrium refinement, the *Laplacian Criterion*, which provides a tractable and behaviorally grounded tool for equilibrium selection. The Principle of Insufficient Reason, already used in both Global Games and Level-k reasoning models, provides a natural foundation for equilibrium selection when agents face uncertainty about others' strategies. Beyond this specific application, the refinement offers a tractable and behaviorally motivated tool for analyzing strategic communication in a broad class of signaling environments.

Specifically, I present a general version of the model in Vaccari (2023a), allowing for one or more informed senders who communicate with an uninformed decision maker choosing between two actions. I focus on the case of conflicting interests between the senders and the receiver. I report the characterization of equilibria in both the monopoly and com-

petitive cases, and I evaluate the receiver's welfare in each environment. After showing that welfare comparisons among the multiple equilibria surviving standard refinement criteria remain inconclusive, I define the *Laplacian Criterion*. Applying this criterion, I identify the class of *Adversarial Equilibria* as the only equilibrium class surviving in competitive environments, and the least informative equilibrium as the only one surviving under monopoly. Extending the analysis to cases with three or more senders increases the complexity of the model. I show that the fully revealing equilibrium survives the *Laplacian Criterion* when there are more than two senders. However, I cannot exclude the existence of other, less informative equilibria that remain robust. This calls for further research to determine whether any such equilibria survive in the case of multiple senders, in order to establish definitively that greater competition always benefits uninformed decision makers.

The remainder of the paper proceeds as follows. Section 2.2 reviews the related literature on costly talk, welfare in strategic communication games, equilibrium refinements in signaling games, and the Principle of Insufficient Reason. Section 2.3 presents the general model, the equilibrium characterization for both monopoly and competition, and some preliminary welfare considerations. Section 2.4 introduces the novel behavioral criterion. Section 2.5 develops the welfare comparison under the *Laplacian Criterion*. Section 2.6 extends the analysis to the case in which there are more than two senders. Section 2.7 concludes.

2.2 Literature

This paper primarily contributes to the growing literature on costly talk, which examines strategic information transmission when sending messages involves real costs. Within this framework, we study how competition among informed senders affects the welfare of an uninformed decision maker. We also make a methodological contribution to the broader theory of equilibrium refinement in signaling games. To address the limitations of existing criteria in comparing welfare across different communication structures, we introduce a new refinement based on the Principle of Insufficient Reason. This principle, which has been invoked in several other contexts in game theory and behavioral economics, provides a natural behavioral foundation for our notion of Laplacian beliefs.

Costly talk. This paper primarily contributes to the literature on costly talk and strategic communication with multiple senders. Classic models such as Battaglini (2002), Krishna and Morgan (2001), and Milgrom and Roberts (1986) show that with costless or verifiable communication, competition among senders leads to full information revelation. In contrast, when misreporting is possible but costly, equilibria may involve only partial revelation. Work in single-sender settings (Ottaviani and Squintani, 2006; Kartik, Ottaviani, and Squintani, 2007; Y. Chen, Kartik, and Sobel, 2008; Kartik, 2009; Y. Chen, 2011) shows that bounded state spaces rule out full revelation and lead to coexistence of revealing and non-revealing equilibria.

Few papers study costly communication with multiple senders. Emons and Fluet (2009) and Kartik, Lee, and Suen (2021) show that competition can increase information transmission under specific conditions. Our paper builds on the theoretical structure by Vaccari, 2023a; Vaccari, 2023b showing that, under the *Laplacian Criterion*, competition always benefits the receiver regardless of biases or reporting costs. In this sense, it bridges results from costly talk and behavioral refinements of equilibrium selection in signaling games.

Welfare considerations in strategic communication games. Classic theories of information transmission suggest that the informational consequences of competition depend on the underlying communication structure. In disclosure environments, rivalry among senders generates stronger incentives for transparency, culminating in full revelation of private information (Gentzkow and Shapiro, 2008). In cheap talk settings, the addition of further senders cannot reduce the receiver's information set, as their messages can be strategically disregarded (Crawford and Sobel, 1982; Milgrom and Roberts, 1986). Taken together, these results imply that, within such canonical models, competition cannot reduce the receiver's welfare relative to monopoly. Recent work has challenged this view by allowing richer communication technologies and testing welfare effects empirically. Kakhbod and Loginova (2023) bridge the hard- and soft-talk literature by letting agents choose between verifiable (hard) and non-verifiable (soft) communication, showing that while verifiability improves information accuracy, it can crowd out cheap talk, making its welfare impact cost-dependent.

Equilibrium concepts in signaling games. The study of strategic communication originates with Spence (1973b) and Crawford and Sobel (1982), who formalized how informed agents convey private information through messages or signals. The predictive power of these models crucially depends on the equilibrium concept employed. Early analyses relied on the Nash Equilibrium (NE), in which each player's strategy is a best response to others' strategies. However, because NE assumes simultaneous moves and perfect information, it is inadequate for dynamic signaling environments. The Bayes–Nash Equilibrium (BNE) extends this idea to settings with incomplete information, requiring each type of each player to maximize expected utility given beliefs derived from the common prior and Bayes' rule.

Dynamic games of incomplete information, such as signaling games, require stronger equilibrium notions. The standard framework is the Perfect Bayesian Equilibrium (PBE) (Fudenberg and Tirole, 1991), which combines optimality (sequential rationality) and belief consistency with

Bayes' rule. A closely related concept is the Sequential Equilibrium (SE) of Kreps and R. Wilson (1982), which refines PBE by ensuring that beliefs are well-defined even at off-path information sets. Further refinements, such as Selten's Perfect Equilibrium (PE) (Selten, 1975) and Grossman and Perry's Perfect Sequential Equilibrium (PSE) (S. J. Grossman and Perry, 1986), strengthen these requirements by imposing credibility and robustness to small perturbations.

While these refinements sharpen predictions within a given signaling structure, they do not necessarily select among multiple equilibria or allow welfare comparisons across environments. This limitation has motivated a large body of work that refines equilibrium beliefs following off-path deviations.

Refinements and belief restrictions. Several criteria have been developed to eliminate implausible equilibria by restricting how receivers update beliefs after deviations. The *Intuitive Criterion* of Cho and Kreps (1987) rules out beliefs assigning positive probability to types that could never gain from a deviation. Building on this, the *Divinity* refinement gives greater weight to types that would benefit more from deviating, and *Universal Divinity* strengthens this further by making the restriction independent of the prior distribution. The *D1* and *D2* criteria impose related conditions: after a deviation, the receiver should place positive probability only on the types with the strongest incentives to deviate, ruling out less motivated types. Finally, the *Never a Weak Best Response* (NWBR) condition of Manelli (1997) eliminates deviations that are merely weakly optimal, thus subsuming many of the previous refinements.

Together, these refinements improve the internal consistency of signaling equilibria, but they remain silent on broader welfare issues, such as whether competition among multiple senders benefits or harms the receiver. This question becomes particularly subtle under costly talk, a communication environment intermediate between cheap talk and verifiable disclosure, where incentives to misreport interact with costly information revelation.

Principle of Insufficient Reason. A complementary line of research addresses how agents form beliefs when objective probabilities are unknown. The Principle of Insufficient Reason, or Principle of Indifference, states that when no reason exists to favor one outcome over another, each should be assigned an equal probability (Gilboa, 2009, Chapter 4). This principle plays a foundational role in behavioral and bounded rationality models of belief formation.

In behavioral game theory, the *Level-k model* of strategic reasoning (Nagel, 1995; Stahl and P. W. Wilson, 1995) provides a direct application of this principle. The Level-0 type behaves nonstrategically, randomizing uniformly across available actions. Since higher-level players have no reason to believe that Level-0 favors any action, they form uniform (Laplacian) beliefs about the base level's behavior. As Battigalli, Charness, and Dufwenberg (2013) observes, this assumption corresponds to the Principle of Insufficient Reason applied to strategic uncertainty.

The same idea appears in the global games literature. In their seminal contribution, Carlsson and Damme (1993) invoke the Principle of Insufficient Reason to justify uniform second-order beliefs before introducing private signals. This assumption provides a behavioral foundation for equilibrium selection, leading to the risk-dominant equilibrium under complete uncertainty.

Our refinement builds on this behavioral logic. We introduce the notion of *Laplacian beliefs*, under which agents, when uncertain about the true probability distribution, assign equal probability to all feasible outcomes. By embedding this belief structure within signaling games with costly communication, we develop a new refinement, the *Laplacian Criterion*, that selects the least informative equilibrium as uniquely stable. This allows us to formally demonstrate that competition among informed senders always benefits the receiver, regardless of sender bias or misreporting technology. Conceptually, our approach bridges the rationalist refinement tradition in game theory with behavioral models of belief formation grounded in the Principle of Insufficient Reason.

2.3 The Model

Setup. There are $N \in \mathbb{N}^+$ senders, indexed by $j \in S = \{1, \ldots, N\}$, and one decision maker (also called receiver, dm). The case $N = 1$ corresponds to a monopoly of information, while $N \geq 2$ corresponds to competition. Nature selects a state θ according to a distribution F with density f and full support $\Theta \subseteq \mathbb{R}$. Each sender j privately observes θ and delivers a report $r_j \in \Theta$ to the decision maker. After observing the report(s) $r = (r_1, \ldots, r_N)$, but not the true state θ, the decision maker chooses an action $a \in \{\ominus, \oplus\}$.

Payoffs. Each player $i \in S \cup \{dm\}$ receives payoff

$$U_i(a, \theta) = \begin{cases} 0, & \text{if } a = \ominus, \\ u_i(\theta), & \text{if } a = \oplus, \end{cases}$$

where $u_i : \Theta \to \mathbb{R}$ is continuous, strictly increasing, and satisfies $u_i(\tau_i) = 0$ for some $\tau_i \in \Theta$. Thus, player i prefers action $\oplus$ to $\ominus$ if and only if $\theta \geq \tau_i$. Figure 5 illustrates preferred actions in the competitive case ($N = 2$, $\tau_1 < 0 < \tau_2$).

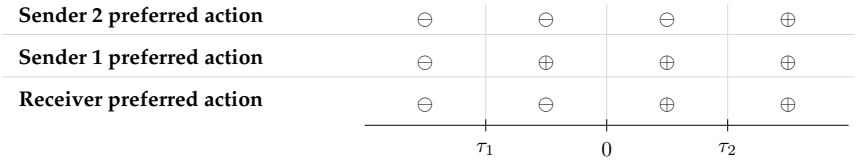


Figure 5: *Preferred actions in competition ($\tau_1 < 0 < \tau_2$).*

Misreporting costs. Each sender $j \in S$ bears a cost $C_j(r_j, \theta)$ when reporting r_j in state θ. The cost function C_j is continuous, strictly increasing in $|r_j - \theta|$, differentiable for $r_j \neq \theta$, and satisfies $C_j(\theta, \theta) = 0$. Thus, truthful reporting is costless, while misreporting becomes more costly the further the report is from the truth. Moreover, for any $r_j \in \Theta$, $C_j(r_j, \theta) > C_j(r_j, \theta')$ if $|r_j - \theta| > |r_j - \theta'|$, meaning distant lies are cheaper when coming from nearby states.

Strategies and beliefs. A pure strategy of sender j is a function $\rho_j : \Theta \to \Theta$, with $\rho_j(\theta)$ denoting the report in state θ. A mixed strategy is a probability distribution $\varphi_j : \Theta \to \Delta(\Theta)$ with support $S_j(\theta)$. The decision maker forms posterior beliefs $p(\theta|r)$, where r is the report (monopoly) or report profile (r_1, r_2) (competition). Given these beliefs, her best response is

$$\beta(r) = \arg \max_{a \in \{\ominus, \oplus\}} \mathbb{E}_p[u_{dm}(a, \theta) \mid r].$$

Solution concept. The solution concept is *Perfect Bayesian Equilibrium* (PBE). A PBE is a profile $(\rho_1, \ldots, \rho_N, p)$ such that each sender's strategy is optimal given others' strategies, beliefs, and the decision maker's rule β, and beliefs p are updated via Bayes' rule whenever possible.

Reach functions. Misreporting costs significantly affect information transmission, as certain reports cannot be profitably delivered from specific states. The notion of reach captures how far a sender can misreport before incurring a sure loss. Since senders may misreport by either exaggerating or understating the realized state, it is necessary to define two sender-specific, state-dependent thresholds—one for each direction of the lie. Reports that exceed these thresholds in a given state are unprofitable.

We define the reach of sender j in state θ as the report whose associated cost exactly offsets j's potential gain. Formally, the reach of sender j in state θ, when $\Delta u_j(0) > 0$, is given by

$$\bar{r}_j(\theta) := \max \{r_j \in \Theta \text{ s.t. } |\Delta u_j(\theta)| = C_j(r_j, \theta)\},$$

while when $\Delta u_j(0) < 0$ it is

$$\underline{r}_j(\theta) := \min \{r_j \in \Theta \text{ s.t. } |\Delta u_j(\theta)| = C_j(r_j, \theta)\}.$$

The reach is computed under the condition that the sender's report is persuasive. Intuitively, in equilibrium, sender j will never deliver reports higher than $\bar{r}_j(\theta)$ or lower than $\underline{r}_j(\theta)$, as these reports are strictly dominated by truthful reporting, independently of the receiver's decision.

We also refer to the *inverse reach* functions, $\bar{r}_j^{-1}(r)$ and $\underline{r}_j^{-1}(r)$, which denote respectively the lowest and highest states from which a sender would be willing to send report r.

2.3.1 One Sender: Equilibria and Welfare

We now explore the equilibrium structure and welfare implications in the case of a single sender ($N = 1$), i.e., the *monopoly* setting. This baseline helps illustrate how costly communication operates in isolation before introducing competition among senders. For simplicity, we report here the case for $\tau_1 > 0$, the case $\tau_1 < 0$ is symmetric.

Equilibrium. Consider first the monopoly case ($N = 1$), where a single informed sender communicates with the receiver. An equilibrium is a pair $(\rho(\theta), p(\theta \mid r))$ satisfying the sender's optimal reporting strategy and the receiver's consistent beliefs.

In the monopolistic setting, there is a continuum of equilibria, ranging from the least to the most informative equilibrium for the receiver. Across all equilibria, the sender truthfully reports extreme states while pooling moderate states near $\tau_{dm} = 0$.

The set of states where the strategy of the sender prescribes to pool characterizes the information transmission of the equilibrium.

In particular, the most informative equilibrium resembles the fully revealing one: the sender 1 ($\tau_1 < 0$) reports the pooled value $\bar{r}_1(0)$ for all states $\theta \in [0, \bar{r}_1(0)]$, and reports truthfully, $\rho_1(\theta) = \theta$, for states outside this interval. In this case, even though the receiver does not know the exact value of the state, she always learns the true sign, letting her efficiently choose her preferred action.

In the other equilibria, the set of states where the sender pools the reports shifts to the right, providing less and less information to the receiver.

Figure 6 illustrates the sender's reporting behavior in the least informative equilibrium, where the monopolistic sender delivers the same pooling report r^*, whenever θ takes values in the zero-centered interval

$[r^*, \bar{r}_1^{-1}(r^*)]$. Otherwise, the sender reports truthfully.

In our analysis, the least informative equilibrium of this type will be particularly central. For this reason, we illustrate it in detail below and graphically in Figure 6. All the derivations are deepened by Vaccari, 2020.

The least informative equilibrium in the monopoly case is such that the following statements hold.

(i) When the sender's bias $\tau_1 > 0$, the reporting strategy is:

$$\rho(\theta, q) = \begin{cases} r^* = \{r \in \Theta \mid \mathbb{E}_f[\theta \mid \theta \in (r, \bar{r}_1^{-1}(r))] = 0\}, & \text{if } \theta \in (r^*, \bar{r}_1^{-1}(r^*)) \\ \theta, & \text{otherwise.} \end{cases}$$

(ii) When $\tau_1 < 0$,

$$\rho(\theta, q) = \begin{cases} r^* = \{r \in \Theta \mid \mathbb{E}_f[\theta \mid \theta \in (\underline{r}_1^{-1}(r), r)] = 0\}, & \text{if } \theta \in (\underline{r}_1^{-1}(r^*), r^*) \\ \theta, & \text{otherwise.} \end{cases}$$

(iii) When $\tau_1 = 0$, the sender reports truthfully: $\rho(\theta) = \theta$ for all $\theta \in \Theta$.

The receiver's beliefs $p(\theta \mid r)$ are derived by Bayes' rule wherever possible and satisfy $\mathbb{E}_p[\theta \mid r^*] = 0$. For off-path reports r, beliefs are assigned as follows:

$$\begin{cases} \mathbb{E}_p[\theta \mid r] < 0, & \text{if } \tau_1 < 0, \\ \mathbb{E}_p[\theta \mid r] > 0, & \text{if } \tau_1 > 0, \\ p(\theta \mid r) \text{ degenerates at } \theta = r, & \text{if } \tau_1 = 0. \end{cases}$$

Consequently, the receiver's decision rule is:

$$\beta(r) = \begin{cases} \oplus, & \text{if } r \geq r^* \text{ and } \tau_1 < 0, \\ \ominus, & \text{if } r \leq r^* \text{ and } \tau_1 > 0, \\ \oplus, & \text{if } r \geq 0 \text{ and } \tau_1 = 0. \end{cases}$$

In equilibrium, the monopolistic sender truthfully reports extreme states, where persuasion is too costly, but pools intermediate states around the receiver's decision threshold τ_{dm}.

All monopoly equilibria are in pure strategies.

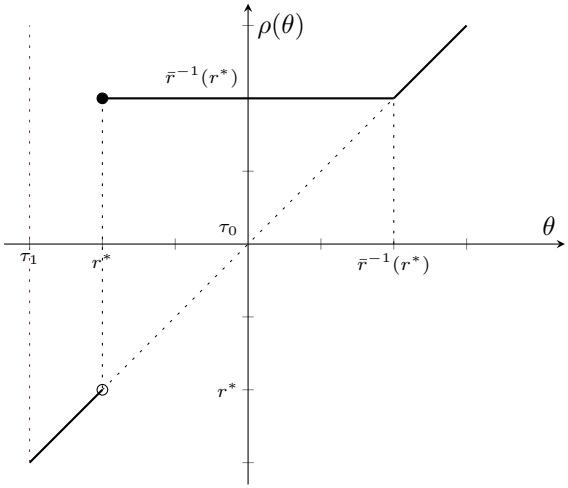


Figure 6: *Graphical representation of an example of equilibrium strategy of the monopoly sender.*

Notes: For all $\theta \leq \tau_1$, the sender prefers $\ominus$. For $\theta \in \{0, \tau_1\}$, there is a conflict of interest, and the sender would like to convince the decision maker that θ is negative. It is profitable for the sender to misreport only for values lower than $\bar{r}_1^{-1}(r^*)$, but they also need to misreport for values lower than τ_0 to convince the decision maker to select $\ominus$. The receiver is indifferent between the two alternatives when observing report r^*.

Receiver's welfare. As shown in the previous section, the monopolistic environment admits a multiplicity of equilibria, each implying a different level of informativeness and, consequently, a different level of welfare for the decision maker.

These equilibria span the entire range from the least to the most informative outcomes, so the receiver's welfare ranges from the level corresponding to no useful information to that of full information.

We can express the receiver's welfare in the monopoly case as the full-information benchmark minus the informational loss induced by persua-

sive communication:

$$W_{dm}^{\text{fi}} = \int_0^{\max \Theta} f(\theta)\, u_{dm}(\theta)\, d\theta,$$

$$W_{dm}^{M} = \int_{\bar{r}_1^{-1}(r^*)}^{\max \Theta} f(\theta)\, u_{dm}(\theta)\, d\theta = W_{dm}^{\text{fi}} + \int_{\bar{r}_1^{-1}(r^*)}^{0} f(\theta)\, u_{dm}(\theta)\, d\theta.$$

The second integral represents the expected welfare loss relative to full information. Different values of r^* correspond to different pooling thresholds, and thus to different welfare levels, from the fully revealing equilibrium (maximum welfare, when $\bar{r}_1^{-1}(r^*) = 0$) to the *least informative equilibrium* (minimum welfare, r^* such that $\mathbb{E}_p[\theta \mid r^*] = 0$), in which communication conveys no useful information and the sender persuades the decision maker in all states.

This multiplicity underscores the need for a more stringent refinement criterion, developed in the next section, to behaviorally rank equilibria and yield a unique welfare prediction for the receiver.

Standard refinements such as the Never a Weak Best Response (NWBR) criterion fail to eliminate any of these equilibria: all monopoly equilibria survive NWBR (see section 2.8). Therefore, additional structure is required to derive meaningful welfare comparisons.

2.3.2 Two Senders: Equilibria and Welfare

We now turn to the case of two informed senders ($N = 2$), which introduces competition in communication. This setting allows us to explore how strategic interaction between biased senders affects the informativeness of equilibrium and the welfare of the decision maker.

Equilibrium. An equilibrium in this environment is a triple (ρ_1, ρ_2, p), where each reporting strategy ρ_j is optimal given the other sender's strategy (ρ_{-j}), the receiver's beliefs p, and her decision rule β, while p satisfies Bayes' rule whenever possible.

Vaccari (2023a) show that among all possible equilibria, the only plausible[1] ones are those that satisfy two conditions on beliefs, collectively defining the class of *Adversarial Equilibrium* (AE).

An *Adversarial Equilibrium (AE)* is an equilibrium that satisfies:

(i) **Strong Monotonicity (sM):** For all conflicting reports $r_1 \geq 0 \geq r_2$ such that $\underline{r}_2(0) < r_2$ and $r_1 < \overline{r}_1(0)$,

$$\frac{d}{dr_j} U_{dm}(r_1, r_2) > 0, \quad j \in \{1, 2\}. \tag{sM}$$

This condition ensures that higher reports by either sender cannot reduce the receiver's expected utility from choosing action R^+. As a result, each sender's strategy is monotone: $r_j \geq \theta$ for $\theta \geq \tau_j$, and $r_j \leq \theta$ for $\theta \leq \tau_j$.

(ii) **Dominance (Dom):** At the boundary reports $(\overline{r}_1(0), \underline{r}_2(0))$ and at the symmetric profile $(0, 0)$, the receiver is indifferent:

$$U_{dm}(\overline{r}_1(0), \underline{r}_2(0)) = U_{dm}(0, 0) = 0. \tag{Dom}$$

This condition implies that when both senders make extreme or symmetric reports, the receiver interprets these messages as evidence that the underlying state is zero.

The (sM) condition provides a general behavioral restriction that applies to a wide class of persuasion and signaling problems (see Vaccari, 2023a; Vida, Honryo, and Azacis, 2022). The (Dom) condition, on the other hand, serves as a normalization ensuring symmetry in the receiver's interpretation of the senders' messages. Together, these two properties identify the class of Adversarial Equilibria (AE).

Vaccari (2023a) provides a complete characterization of AE. For the sake of this contribution, it is important to know that in a symmetric environment[2], both senders always report truthfully when the realized

[1] An equilibrium is plausible when it is $\epsilon - robust$ Battaglini, 2002 and supported by unprejudiced beliefs(K. Bagwell and Ramey, 1991).

[2] Linear utilities, quadratic costs, and symmetric priors, see Vaccari, 2023a

state lies outside the truthful cutoffs, defined as θ_l and θ_h. For values between θ_l and zero, the sender 1 tells the truth with some probability $\alpha_1 \in (0, 1)$. For θ between zero and θ_h, then the sender 2 is reporting the truth with some probability α_2.

Receiver welfare. Computing welfare under competition is generally complex, but for our purposes, it is sufficient to provide informative welfare bounds around the Adversarial Equilibrium. Let W_{dm}^{AE} denote the receiver's equilibrium welfare under AE. We can establish that it lies between two natural benchmarks.

- *Lower bound* ($\underline{w}_{dm}$): The receiver ignores sender 2 and relies solely on sender 1's message. Formally, she chooses $\oplus$ whenever $r_1 \geq 0$ and $\ominus$ otherwise. Let θ_l denote the cutoff state at which the receiver is exactly indifferent between the two actions, that is the lower state in which sender 1 misreports with positive probability. For each sender j, let $\alpha_j(\theta) \in [0, 1]$ be the probability that the sender reports truthfully in state θ, and $(1 - \alpha_j(\theta))$ the probability of misreporting. These probabilities are determined by the mixed strategies reported in Vaccari (2023a). Since relying on a single sender ignores valuable information, this strategy is not sequentially rational and therefore must yield weakly lower welfare than in equilibrium:

$$\underline{w}_{dm} = \int_0^{\max \Theta} f(\theta)\, u_{dm}(\theta)\, d\theta + \int_{\theta_l}^{0} f(\theta)\, u_{dm}(\theta)\, (1 - \alpha_1(\theta))\, d\theta.$$

  The first term represents welfare when the receiver's decision is always correct for $\theta \geq 0$, while the second captures expected losses from possible mistaken choices when $\theta \in \{\theta_l, 0\}$, which occur whenever sender 1 misreports.

- *Upper bound* ($\bar{w}_{dm}$): The receiver again makes no mistakes when $\theta > 0$, but now errs less frequently when $\theta < 0$. Suppose that errors arise only when sender 1 misreports while sender 2 is truthful, so that the receiver correctly interprets disagreement as a signal of

negative states (not possible in AE). The associated welfare is

$$\bar{w}_{dm} = \int_0^{\max\Theta} f(\theta)\, u_{dm}(\theta)\, d\theta + \int_{\theta_l}^{0} f(\theta)\, u_{dm}(\theta)\, (1-\alpha_1(\theta))\, \alpha_2(\theta)\, d\theta.$$

This bound is necessarily higher, since the probability of a mistaken choice for $\theta < 0$ is now reduced by the factor $\alpha_2(\theta)$, the likelihood that sender 2 truthfully reveals the negative state when sender 1 is misreporting.

Hence, equilibrium welfare satisfies:

$$\underline{w}_{dm} < W^*_{dm} < \bar{w}_{dm}.$$

Both bounds share the full-information welfare as their first term, while the second term captures the informational inefficiency due to strategic misreporting. The gap between $\underline{w}_{dm}$ and $\bar{w}_{dm}$ quantifies the residual uncertainty that competition cannot fully eliminate.

2.3.3 Preliminary Welfare Considerations

Comparing the decision maker's welfare across monopoly and competition, even when restricting attention to monotone frameworks, leads to inconclusive results. The source of this ambiguity lies in the multiplicity of equilibria that survive standard equilibrium refinements, as shown in Section 2.8. Within the set of monopoly equilibria that satisfy standard refinements, there exist some that yield higher welfare for the receiver than certain equilibria in competition.

Formally, in the monopoly case, the upper bound of attainable welfare is:

$$W^M_{dm} = \int_{\bar{r}_1^{-1}(r^*)}^{\max\Theta} f(\theta) u_{dm}(\theta)\, d\theta = W^{\text{fi}}_{dm} + \int_{\bar{r}_1^{-1}(r^*)}^{0} f(\theta) u_{dm}(\theta)\, d\theta,$$

with r^* such that it equates to the welfare of full revelation.

In contrast, under competition, the receiver's welfare is bounded below by:

$$\underline{w}_{dm} = \int_0^{\max\Theta} f(\theta)\, u_{dm}(\theta)\, d\theta + \int_{\theta_l}^{0} f(\theta)\, u_{dm}(\theta)\, (1-\alpha_1(\theta))\, d\theta.$$

Hence, while some competitive equilibria yield higher welfare than most monopolistic ones, others are dominated by more informative monopoly equilibria. From a theoretical standpoint, the existing refinement criteria do not allow us to rank these outcomes conclusively, as discussed in section 2.8.

This ambiguity calls for a sharper selection principle. In particular, we seek a refinement that (i) preserves the behavioral plausibility of standard refinements, (ii) remains applicable across both monopoly and competition, and (iii) provides unambiguous welfare predictions. To achieve this, we introduce in the next section a behavioral refinement based on Laplace's Principle of Insufficient Reason, which yields sharper equilibrium selections in both environments and restores a clear welfare ranking.

Standard equilibrium refinements are not sufficient to resolve this ambiguity. Criteria such as the Intuitive Criterion, Divinity, or NWBR restrict beliefs by ruling out implausible deviations based on payoff dominance. However, in our setting, these refinements do not rule out equilibria sustained by off-path beliefs that require the receiver to assign disproportionate weight to some feasible states without clear justification. The Laplacian Criterion excludes such equilibria by imposing that, when several states remain equally feasible after a deviation, the receiver applies the Principle of Insufficient Reason and assigns them equal probability. As a result, both relatively informative and uninformative equilibria may survive traditional criteria, leaving welfare comparisons across monopoly and competition inconclusive. This limitation highlights the need for a refinement that disciplines belief formation more tightly in the presence of residual strategic uncertainty, which motivates the introduction of the Laplacian Criterion.

2.4 The Laplacian Criterion

Analogous to the Intuitive Criterion of Cho and Kreps, 1987, the *Laplacian Criterion* provides a systematic test for assessing the plausibility of a given equilibrium outcome in signaling games. It builds on the intuition that when an off-path report is observed, the receiver should only form beliefs over those states in which such a deviation could, in principle, be profitable for the sender. Beliefs assigning positive probability to implausible states are thus ruled out. The Intuitive Criterion is not directly applicable to settings with more than one sender. We expand the concept to this enriched framework, adding a behavioral extension based on the Principle of Insufficient Reason.

Let Θ denote the state space. For each report profile $\mathbf{r} := \{r_1, ..., r_N\}$, with $N \in \mathbb{N}^+$, and any posterior belief $p(\theta|r_1, ..., r_N)$, such that

$$\int_{\Theta} p(\theta|r_1, ..., r_N)\, d\theta = 1,$$

the receiver's best-response correspondence is defined as

$$BR_{\forall p}(\theta, r_1, ..., r_N) =$$

$$\bigcup_{p:\int_{\Theta} p(\theta|r_1,...,r_N)=1} \arg\max_{a} \int_{\Theta} p(\theta|r_1, ..., r_N)\, u_{r_1,...,r_N}(a, \theta)\, d\theta.$$

That is, $BR_{\forall p}(\theta, r_1, ..., r_N)$ collects all receiver actions that could be optimal under some posterior belief about θ, given reports $r_1, ..., r_N$.

Fix an equilibrium outcome and denote by $u_j^*(\theta)$ the sender's expected equilibrium payoff in state θ. The first step in applying the *Laplacian Criterion* is to construct the set of states in which deviation to report r_j could be profitable for the sender $j \in S$, under some belief:

$$K(r_j) = \Big\{\theta \in \Theta \;\Big|\; u_j^*(\theta) < \max_{a \in BR_{\forall p}(\theta, r_1^*, ..., r_j, ..., r_N^*)} u_j(r, a, \theta)\Big\}.$$

The set $K(r_j)$, therefore, contains the subset of states where the equilibrium strategy does not strictly dominate deviation to r_j.

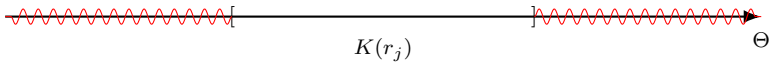


Figure 7: *The subset $K(r_j)$ of states which can somehow justify a deviation from the equilibrium, under some beliefs.*

Once the credible set $K(r_j)$ is determined, the next step concerns the receiver's beliefs. Any posterior assigning positive probability to states outside $K(r_j)$ after observing reports $\{r_1^*, \ldots, r_j, \ldots, r_N^*\}$ is not credible. Within $K(r_j)$, however, the receiver has no reason to consider any particular state more likely than another. By the Principle of Insufficient Reason, the receiver therefore assigns a uniform distribution over the set $K(r_j)$. Formally, the posterior density is given by $p_K(\theta|r_j)$:

$$p_K(\theta|r_j) = \begin{cases} \frac{1}{\mu(K(r_j))} & \text{if } \theta \in K(r_j) \\ 0 & \text{if } \theta \notin K(r_j) \end{cases}$$

where $\mu(K(r_j))$ denotes the length (Lebesgue measure) of the interval $K(r_j)$. That is, the receiver's posterior is uniform on the convex set $K(r_j)$, assigning equal density to all states within it.

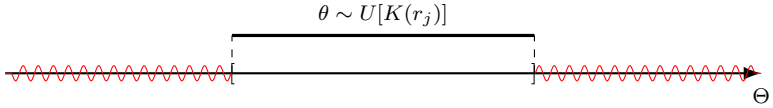


Figure 8: *Laplacian beliefs assign uniform probability over $K(r_j)$, the subset of states where deviation to r could be profitable.*

Accordingly, the receiver's best response under these *Laplacian beliefs* is given by

$$BR_{p_K}(K, r_1^*, ..., r_j, ..., r_N^*) = \arg\max_a \int_K p_K(\theta|r_j)\, u_{r_1^*,\ldots,r_j,\ldots,r_N^*}(a, \theta)\, d\theta.$$

Finally, the equilibrium fails the *Laplacian Criterion* if there exists a state θ' and an off-path report r_j' such that

$$u_j^*(\theta') < \min_{a \in BR_{p_K}(K(r_j'), r_j')} u_j(r_j', a, \theta').$$

That is, the equilibrium is eliminated whenever a sender j could profitably deviate to r'_j while anticipating that the receiver would hold Laplacian beliefs, uniform over the credible set $K(r'_j)$, upon observing that deviation.

The last step of our *Laplacian Criterion* consists of restricting the attention to Strongly Monotonic Games. This is in line with the approach by Vaccari, 2023a and Vida, Honryo, and Azacis, 2022. Strong Monotonicity (sM) in the setting of $N \in \mathbb{N}^+$ senders is presented below, case by case.

- $N = 1$.

$$\frac{d}{dr}U_{dm}(r) > 0, \qquad \text{(sM, } N = 1\text{)}$$

- $N = 2$. For all conflicting reports $r_1 \geq 0 \geq r_2$ such that $\underline{r}_2(0) < r_2$ and $r_1 < \overline{r}_1(0)$,

$$\frac{d}{dr_j}U_{dm}(r_1, r_2) > 0, \quad j \in \{1, 2\}. \qquad \text{(sM, } N = 2\text{)}$$

  This condition ensures that higher reports by either sender cannot reduce the receiver's expected utility from choosing action R^+. As a result, each sender's strategy is monotone: $r_j \geq \theta$ for $\theta \geq \tau_j$, and $r_j \leq \theta$ for $\theta \leq \tau_j$.

- $N \geq 3$.

$$\text{If } r_1 = r_2 = \cdots = r_N \implies \frac{\partial U_{dm}(r_1, \ldots, r_N)}{\partial r_j} = 0, \quad j \in S,$$

$$\text{otherwise} \quad \frac{\partial U_{dm}(r_1, \ldots, r_N)}{\partial r_j} > 0, \quad j \in S.$$

(sM, $N \geq 3$)

  Intuitively, when several senders communicate simultaneously, the influence of any single sender on the receiver's posterior is diluted. If one sender deviates while the remaining $N - 1$ reports converge to the same message consistent with a particular state, the receiver's beliefs remain effectively pinned to that state. As a result, unilateral deviations cannot profitably affect the receiver's action, reinforcing truthful and monotone reporting as the only robust behavior under competition.

The *Laplacian Criterion* thus refines equilibrium selection by eliminating outcomes that depend on unreasonable off-path beliefs, for any number of senders. It parallels the Intuitive Criterion of Cho and Kreps, 1987 but departs from it by relying on the Principle of Insufficient Reason rather than on dominance arguments. Furthermore, unlike the Intuitive Criterion, this approach can be applied under both monopoly and competitive conditions, enabling a consistent comparison of equilibria that survive the test.

2.4.1 Monopoly ($N = 1$)

Lemma 1. *The unique monopoly equilibrium surviving the* Laplacian Criterion *is the least informative equilibrium.*

Proof. Consider the one-sender case. Suppose $\tau_1 < 0$ (the case $\tau_1 > 0$ is symmetric). Let r^* denote the equilibrium threshold such that the receiver chooses $a = \oplus$ if and only if $r \geq r^*$. In equilibria where $\mathbb{E}_p[\theta \mid r^*] > 0$, the sender can profitably deviate.

Indeed, take an off-path report $r' < \bar{r}_1^{-1}(r^*)$. Under Laplacian beliefs, the receiver assigns equal probability to all states consistent with r'. Since the feasible state set given r' lies weakly below $\bar{r}_1^{-1}(r^*)$, there always exists r' sufficiently close to $\bar{r}_1^{-1}(r^*)$ such that

$$\mathbb{E}_p[\theta \mid r'] \geq 0,$$

which induces the favorable action $a = \oplus$. Because r' is less extreme than $\bar{r}_1^{-1}(r^*)$, the deviation is strictly cheaper while preserving the desired action, contradicting the equilibrium.

Therefore, equilibria in which $\mathbb{E}_p[\theta \mid r^*] \neq 0$ cannot survive the Laplacian refinement. Only equilibria satisfying $\mathbb{E}_p[\theta \mid r^*] = 0$, where the receiver is exactly indifferent between actions at the pivotal report, are plausible. Among these least informative equilibria, some also satisfy Strong Monotonicity (sM); we exclude those sustained by beliefs that do not update positively in response to a higher signal. These correspond precisely to the least informative equilibria, in which pooling around the decision threshold is maximal. □

The *Laplacian Criterion* rules out equilibria supported by implausible off-path beliefs that favor one action over another without evidence. In

monopoly, such beliefs are necessary to sustain more informative equilibria. Under Laplacian beliefs, the receiver assigns equal probability to states consistent with any report, forcing indifference at the threshold. This makes persuasion unprofitable except in the least informative equilibrium, where the sender pools most heavily around the receiver's decision boundary.

2.4.2 Competition ($N = 2$)

Lemma 2. *Let an equilibrium be given in which sender strategies are monotone (sM). If the equilibrium survives the* Laplacian Criterion*, then the Dominance condition (Dom) holds.*

Proof. Recall Dominance condition states that at the boundary reports $(\bar{r}_1(0), \underline{r}_2(0))$ and at the symmetric profile $(0, 0)$, the receiver is indifferent:

$$U_{dm}(\bar{r}_1(0), \underline{r}_2(0)) = U_{dm}(0, 0) = 0.$$

Consider first the report profile $\mathbf{r} = (\underline{r}_2(0), \bar{r}_1(0))$. By definition of the reporting bounds, this profile can only be observed when the true state is $\theta = 0$, otherwise, one of the two is not best replying. Hence, the decision maker's posterior assigns probability one to $\theta = 0$, and $U_{dm}(\mathbf{r}) = 0$.

Now consider the symmetric profile $\mathbf{r} = (0, 0)$. Given monotone strategies (if prefer $\oplus$, never report lower than $theta$, this is implied by strong monotonicity), we have $r_1 \geq \theta$ for all $\theta \geq \tau_1$ and $r_2 \leq \theta$ for all $\theta \leq \tau_2$. Thus, the set of feasible states compatible with $\mathbf{r} = (0, 0)$ reduces again to $\{\theta = 0\}$. By the *Laplacian Criterion*, when agents are indifferent over multiple states consistent with a signal, they assign equal probabilities to them. Since only $\theta = 0$ remains feasible, the posterior collapses to $\theta = 0$, and therefore $U_{dm}(0, 0) = 0$.

Because both $\mathbf{r} = (\underline{r}_2(0), \bar{r}_1(0))$ and $\mathbf{r} = (0, 0)$ yield the same expected utility for the decision maker, the Dominance condition (Dom) holds. □

The *Laplacian Criterion* refines equilibrium beliefs by imposing indifference among equally plausible states. When both senders' reports symmetrically balance around zero, either at the boundaries or at $(0, 0)$, the receiver, lacking further evidence, treats the state as equally likely to be positive or negative. Under this symmetry, the only belief consistent with equal weighting is that the true state is exactly zero. Hence, the decision

maker remains indifferent between actions at those report profiles, which directly corresponds to the Dominance condition.

This implies that if an equilibrium survives the *Laplacian Criterion* in the case $N = 2$, then it belongs to the class of Adversarial.

2.5 Welfare Comparison

Since Section 2.4 established that the only plausible equilibrium under the *Laplacian Criterion* is the least informative pooling equilibrium in monopoly, the Adversarial Equilibrium in two-sender competition, we can now directly compare the receiver's welfare across these cases.

This comparison formalizes the key insight of the paper: as competition among informed senders intensifies, the receiver benefits from progressively more informative communication. Even partial adversarial pressure (with two senders) improves welfare relative to monopoly.

Proposition 1 makes this intuition precise by ranking receiver welfare of monopoly and competition.

Proposition 1. *In symmetric environments (linear utilities, quadratic costs, and symmetric priors, see Vaccari, 2023a), the receiver's welfare increases when moving from monopoly ($N = 1$) to competition with two informed senders ($N = 2$). Specifically:*

$$W_{dm}^{M} \;<\; W_{dm}^{AE},$$

where:

- W_{dm}^{M} *denotes welfare in the least informative monopoly equilibrium (the only one surviving the* Laplacian Criterion*);*
- W_{dm}^{AE} *denotes welfare in the* Adversarial Equilibrium *arising under symmetric competition ($N = 2$).*

Thus, the introduction of a second sender makes communication more informative and improves receiver welfare relative to monopoly.

Proof. We compare the receiver's expected utility under the monopoly and two-sender equilibria that survive the Laplacian refinement.

Step 1. Monopoly ($N = 1$). Lemma 1 shows that, under the *Laplacian Criterion*, the only surviving equilibrium is the least informative pooling one. Denoting by $r^*(0)$ the pooling report, the receiver's welfare is:

$$W_{dm}^{M} = W_{dm}^{\text{fi}} + \int_{\bar{r}_1^{-1}(r^*(0))}^{0} f(\theta)\, u_{dm}(\theta)\, d\theta.$$

which is strictly below the full-information benchmark

$$W_{dm}^{\text{fi}} = \int_0^{\max\Theta} f(\theta)\, u_{dm}(\theta)\, d\theta.$$

Step 2. Competition ($N = 2$). When two senders compete, each tends to counterbalance the other's exaggeration. Lemma 2 shows that, under the Laplacian refinement, only equilibria satisfying *Dominance* and *strong Monotonicity* survive, giving rise to the *Adversarial Equilibrium* (AE).

Lemma 2.3 shows that the receiver's welfare in the case of two senders is bounded below by

$$\underline{w}_{dm} = W_{dm}^{\text{fi}} + \int_{\theta_l}^{0} f(\theta)\, u_{dm}(\theta)\, (1 - \alpha_1(\theta))\, d\theta.$$

Step 3. Comparison. We want to show that

$$|\int_{\bar{r}_1^{-1}(r^*)}^{0} f(\theta)\, u_{dm}(\theta)\, d\theta| > |\int_{\theta_l}^{0} f(\theta)\, u_{dm}(\theta)\, (1 - \alpha_1(\theta))\, d\theta|$$

But since, by Vaccari (2023a), $\alpha \in (0, 1)$. We just need to show that

$$|\bar{r}_1^{-1}(r^*(0))| \geq |\theta_l|.$$

By definition, the cutoff θ_l because of the symmetric environment, is such that $-\theta_l = \underline{r}_2^{-1}(\theta_l)$.

Moreover, again by the symmetric environment, we have that the reach of sender 1 is the same as the inverse reach of sender 2. Therefore $\bar{r}_1(\theta_l) = -\theta_l$.

By the definition, the pair of reports $(\theta_l, -\theta_l)$ should make the decision maker indifferent in a symmetric environment, i.e., $U_{dm}(-\theta_l, \theta_l) = 0$.

The same happens in the least informative monopolistic equilibrium when sender 1 delivers $r^*(0)$: the receiver is indifferent, and $U_{dm}(r^*(0)) = 0$.

Since pooling occurs between $r^*(0)$ and $\bar{r}_1^{-1}(r^*(0))$, it has to be that

$$\bar{r}_1^{-1}(r^*(0)) = \theta_l \implies |\bar{r}_1^{-1}(r^*(0))| \geq |\theta_l|.$$

Hence the lower bound of the welfare in competition is strictly higher than the welfare in the monopoly equilibrium, the following inequality holds true. □

2.6 Three or More Senders

As a natural extension of the main result, we examine how equilibrium outcomes evolve as competition intensifies through the addition of more senders. In particular, we consider the case where $N \geq 3$, so that the decision maker receives reports from multiple informed senders with potentially conflicting biases.

2.6.1 Equilibria and Welfare

Equilibrium. Intuitively, as competition intensifies, the persuasive power of any individual sender diminishes: each sender knows that any exaggeration will be exposed by at least one opposing sender, making truthful communication the only credible strategy. Indeed, this framework admits *Fully Revealing Equilibria* (FRE). In a FRE, all senders report truthfully, $r_j = \theta$ for all $j \in \{1, 2, \ldots, N\}$, and the decision maker's posterior beliefs perfectly coincide with the true state. The fully revealing equilibrium in truthful strategies is the only receiver-efficient equilibrium satisfying generalized versions of (Dom) and (sM), as shown by Vaccari (2023a). Intuitively, when all reports coincide, no sender can profitably deviate, since unilateral misreporting has no impact on the receiver's belief but incurs a cost. If any sender deviates, the receiver detects inconsistency across reports, updates her belief toward the truthful cluster, and disregards the outlier.

However, we still cannot rule out the existence of equilibria involving partial persuasion that may survive the *Laplacian Criterion*. The investigation of such possibilities is left for future research.

Receiver's welfare. Under full revelation, the decision maker always takes the correct action, and welfare attains the first-best benchmark:

$$W_{dm}^{FRE} = W_{dm}^{fi} = \int_{\Theta} f(\theta)\, u_{dm}(\theta)\, d\theta.$$

Hence, when $N \geq 3$, equilibrium communication under costly and verifiable signaling may achieve the first-best outcome.

Nevertheless, we cannot rule out the existence of other equilibria, not analyzed in this study, that may lead to lower welfare.

Laplacian Criterion ($N \geq 3$) We now examine whether the FRE in the case with $N \geq 3$ senders survives the *Laplacian Criterion*.

Lemma 3. *The Fully Revealing Equilibrium (FRE) survives the* Laplacian Criterion *when* $N \geq 3$.

Proof. In a Fully Revealing Equilibrium, no sender finds it profitable to deviate from truthful reporting. When $N \geq 3$, a unilateral deviation by one sender has a negligible impact on the receiver's posterior, since beliefs are determined jointly by multiple independent messages. Consequently, a single deviation cannot shift the receiver's decision, and truthful reporting remains optimal for all senders. These types of equilibria also respect (sM), since a single deviation cannot move the receiver's beliefs, but just a coalitional deviation. Hence, the FRE is robust under the *Laplacian Criterion*. □

Remark. When competition intensifies further ($N \geq 3$), the *Fully Revealing Equilibrium* (FRE) survives the *Laplacian Criterion*, as shown in Lemma 3. In this case, the welfare comparison is straightforward:

$$W_{dm}^{AE} \leq W_{dm}^{FRE} = W_{dm}^{fi}.$$

Hence, receiver welfare may increase with the number of informed senders, approaching the full-information benchmark as N grows.

However, the existence of other equilibria for $N \geq 3$ cannot be ruled out. Some of these equilibria may involve partial revelation or persuasive behavior, potentially leading to welfare levels strictly below the first-best.

2.7 Discussion

This paper investigates whether competition among informed senders benefits an uninformed decision maker in environments where communication is costly and partially verifiable. By bridging the gap between the classic cheap talk and verifiable disclosure frameworks, we show that competition enhances the receiver's welfare.

Methodologically, we introduce a behavioral refinement, the *Laplacian Criterion*, grounded in the Principle of Insufficient Reason. This criterion restricts the set of plausible equilibria in settings with one or two senders, ensuring that welfare in the latter case is always higher than in the former.

The *Laplacian Criterion* allows us to overcome the equilibrium multiplicity that makes welfare comparisons inconclusive under traditional refinements. In particular, we identify the Adversarial Equilibrium as the unique competitive outcome and the least informative equilibrium as the only surviving equilibrium under monopoly. The welfare comparison between these equilibria provides a formal and robust argument that competition among informed senders benefits the receiver.

Beyond its theoretical contribution, the framework speaks to a broad set of real-world environments in which several informed parties attempt to influence a common decision maker while facing some cost of exaggeration or misreporting. One natural example is the case of competing policy advisors, introduced in the Introduction. Advisors may have different preferred policies and therefore an incentive to overstate the severity or desirability of a given state of the world. At the same time, exaggeration is not free, because statements that are too far from the truth may damage credibility or reduce persuasive power. In such settings, the model suggests that competition among advisors can improve the information available to the decision maker, precisely because each advisor anticipates that excessive distortion is more likely to be exposed by the others.

A similar logic may apply in other contexts, such as expert advice, political communication, and media markets. Competing experts may

strategically frame evidence in order to influence a regulator, a firm, or a voter. Politicians and parties may selectively emphasize facts that support their preferred policies, but face reputational or electoral costs if claims are later revealed to be too misleading. Media outlets may slant coverage, yet still be constrained by the risk of losing credibility when rival outlets reveal omitted information or contradictory evidence. In all these cases, communication is neither pure cheap talk nor full verifiable disclosure: messages can be strategically distorted, but distortion is disciplined by some cost. The main result of the paper is therefore especially relevant for environments in which strategic communication is biased but not unconstrained.

This mechanism also helps clarify how the present results differ from those of traditional cheap-talk or persuasion models. In cheap talk, competition cannot hurt the receiver because messages can always be ignored, but this also means that competition disciplines communication only weakly unless additional structure is imposed. In verifiable disclosure and standard persuasion models, by contrast, competition often pushes directly toward full revelation. In the present framework, the effect of competition is more subtle. Competition improves outcomes not because communication is costless or fully verifiable, but because strategic rivalry interacts with costly misreporting: each sender remains biased, yet the presence of competing senders reduces the scope for unilateral exaggeration. This is the central economic mechanism behind the welfare comparison.

Several directions for future research emerge. First, it would be valuable to study in greater depth how these results extend to environments with more than two senders. Second, the *Laplacian Criterion* could be applied to other signaling or strategic communication settings to refine equilibria where traditional approaches fail. Finally, experimental validation could assess both the predictive power of the *Laplacian Criterion* and the welfare effects of competition in laboratory or field contexts.

In sum, this paper provides a clear theoretical foundation for understanding how competition in information provision affects uninformed decision makers and offers a novel, behaviorally grounded tool for equi-

librium selection in signaling games.

2.8 Appendix A: Standard Refinements Fail

Misreporting costs give rise to multiple equilibria, and in some cases, equilibria under monopoly are more informative than those under competition. In this section, we show that standard equilibrium refinements fail to eliminate this multiplicity because they cannot rule out the most informative equilibria in the monopoly case. Manelli (1997) shows that equilibria satisfying the Never-a-Weak-Best-Response (NWBR) test also satisfy several other well-known refinements, such as the Intuitive Criterion and Universal Divinity. Hence, by demonstrating that our equilibria satisfy NWBR, we can extend the result to a broad class of refinements.

NWBR equilibria in monopoly The Never a Weak Best Response (NWBR) criterion eliminates type–report pairs that can never arise as weak best responses. It is the strongest refinement in signaling games: if an equilibrium survives NWBR, then it also survives all weaker refinements (Manelli, 1997). However, in our costly talk setting, NWBR alone is not sufficient to establish that competition yields higher welfare than monopoly. Hence, we need the new *Laplacian Criterion*.

Formally, for each type θ, report r, and belief set T, define:

- $BR(T, r)$: the set of receiver best responses to r when beliefs are concentrated in T;
- $D(\theta, T, r)$: the set of receiver mixed best responses that make type θ strictly prefer r over its equilibrium strategy;
- $D^0(\theta, T, r)$: the set of receiver mixed best responses that make type θ indifferent.

A type–report pair (θ, r) is eliminated under NWBR if

$$D^0(\theta, \Theta, r) \subset \bigcup_{\theta' \neq \theta} D(\theta', \Theta, r).$$

Lemma 4. *All monopolistic equilibria are robust to NWBR.*

Proof sketch. Consider an off-path report r_1. For each possible type θ, compute the set of receiver best responses that make θ indifferent or strictly prefer r_1. These sets collapse to singletons, and as θ varies, they connect smoothly. The key property is that for every θ except one cutoff type, the receiver's best responses that rationalize r_1 are already covered by other types' profitable deviations. Therefore, all type–report pairs are eliminated except the cutoff type, and deviations are unprofitable. Hence, all monopolistic equilibria survive NWBR. □

2.9 Appendix B: Comparative Statics

In this section, we study how the equilibrium cutoff θ_l, the point where sender 1's upper reach meets the swing report, responds to changes in the bias (τ_2) and the misreporting cost (k_2) of the opposing sender. This analysis clarifies how competition affects information transmission and, consequently, welfare.

In competition, a *swing state* θ^s is defined as the point where the decision maker is exactly indifferent between taking actions $\ominus$ and $\oplus$ given the pair of reports (r_1, r_2):

$$U_{dm}(r_1(\theta^s), r_2(\theta^s)) = 0.$$

In symmetric environments, such as when $\tau_1 = -\tau_2$ and $k_1 = k_2$, this implies

$$r_2(\theta^s) = -r_1(\theta^s) \quad \text{and} \quad \theta^s = 0.$$

The swing state thus provides a natural benchmark to assess how changes in senders' preferences or costs shift the receiver's indifference point.

The key goal is to show that, in an Adversarial Equilibrium, the truthful cutoff θ_l does not decrease when τ_2 increases or when k_2 decreases. This ensures that the receiver's lower-bound welfare $\underline{w}_{dm}$ is always at least as high as in the least informative monopolistic equilibrium.

Formally, the cutoff is defined by the condition

$$G_\theta(\theta_l; \tau_2, k_2) = s(\theta_l) - \bar{r}_1(\theta_l) = 0,$$

where $s(\cdot)$ is the swing report and $\bar{r}_1(\cdot)$ is sender 1's upper reach.

Applying the Implicit Function Theorem gives:

$$\frac{d\theta_l}{d\tau_2} = -\frac{\frac{\partial G_\theta}{\partial \tau_2}}{\frac{\partial G_\theta}{\partial \theta_l}}, \qquad \frac{d\theta_l}{dk_2} = -\frac{\frac{\partial G_\theta}{\partial k_2}}{\frac{\partial G_\theta}{\partial \theta_l}},$$

where the partial derivatives are

$$\frac{\partial G_\theta}{\partial \tau_2} = \frac{\partial s(\theta_l)}{\partial \tau_2} - \frac{\partial \bar{r}_1(\theta_l)}{\partial \tau_2}, \qquad \frac{\partial G_\theta}{\partial k_2} = \frac{\partial s(\theta_l)}{\partial k_2} - \frac{\partial \bar{r}_1(\theta_l)}{\partial k_2}.$$

Since $\bar{r}_1$ is independent of τ_2 and k_2, the comparative statics hinge entirely on the sign of the swing report derivatives $\partial s(\theta_l)/\partial \tau_2$ and $\partial s(\theta_l)/\partial k_2$.

2.9.1 Comparative Statics on the Swing (Higher Bias)

To study $\frac{ds(r_1)}{d\tau_2}$, we adopt linear–quadratic primitives for all $i \in \{1, 2, dm\}$ and $j \in \{1, 2\}$:

$$u_i(\theta) = \theta - \tau_i, \quad C_j(r_j, \theta) = (r_j - \theta)^2. \tag{LQ}$$

The senders' reach functions are then

$$\bar{r}_1(\theta) = \theta + \sqrt{\frac{\theta - \tau_1}{k_1}}, \qquad \underline{r}_2(\theta) = \theta - \sqrt{\frac{\tau_2 - \theta}{k_2}}.$$

Their inverses are given by

$$\bar{r}_1^{-1}(r_1) = r_1 + \frac{1 - \sqrt{1 + 4k_1(r_1 - \tau_1)}}{2k_1},$$

$$\underline{r}_2^{-1}(r_2) = r_2 - \frac{1 - \sqrt{1 + 4k_2(\tau_2 - r_2)}}{2k_2}.$$

The swing report $s(r_1)$ is implicitly determined by the equilibrium condition

$$G_s(r_1, s(r_1)) = \int_{\mathcal{I}(r_1, s(r_1))} \theta f(\theta) \frac{(s(r_1) - \theta)(r_1 - \theta)}{(\theta - \tau_1)(\theta - \tau_2)}\, d\theta = 0.$$

Differentiating G_s with respect to $s(r_1)$ yields:

$$\frac{dG_s(\cdot)}{ds(r_1)} = \int_{\mathcal{I}(r_1, s(r_1))} \frac{\theta f(\theta)(r_1 - \theta)}{(\theta - \tau_1)(\theta - \tau_2)}\, d\theta$$

$$+ \mathbb{1}\{\underline{r}_2^{-1}(s(r_1)) < r_1\}\, T(\underline{r}_2^{-1}(s(r_1)), r_1, s(r_1)) \frac{d\underline{r}_2^{-1}(s(r_1))}{ds(r_1)},$$

where

$$T(\theta, r_1, r_2) = \frac{4\theta f(\theta)(r_2 - \theta)(r_1 - \theta)}{(\theta - \tau_1)(\theta - \tau_2)} > 0$$

and

$$\frac{d\underline{r}_2^{-1}(r_2)}{dr_2} = 1 - \frac{1}{\sqrt{1 + 4k_2(\tau_2 - r_2)}} > 0.$$

Hence, $\frac{dG_s(\cdot)}{ds(r_1)} > 0$, ensuring the swing function is locally increasing.

To analyze $\frac{dG_s(\cdot)}{d\tau_2}$, we apply the Leibniz rule:

$$\frac{dG_s(r_1, s(r_1))}{d\tau_2} =$$

$$\int_{\mathcal{I}(r_1, s(r_1))} \theta f(\theta) \frac{(r_2 - \theta)(r_1 - \theta)}{(\theta - \tau_1)(\theta - \tau_2)^2} d\theta$$

$$+ \mathbb{1}\left\{\underline{r}_2^{-1}(s(r_1)) < r_1\right\} T(\underline{r}_2^{-1}(s(r_1)), \cdot) \frac{d\underline{r}_2^{-1}(s(r_1))}{d\tau_2}.$$

Given that $d\underline{r}_2^{-1}(s(r_1))/d\tau_2 > 0$, the second term is positive. The sign of the integral term is typically negative, implying $\frac{dG_s(\cdot)}{d\tau_2} < 0$ in most relevant regions. Therefore, by the IFT, $\frac{ds(r_1)}{d\tau_2} > 0$: a more biased sender 2 raises the swing report.

Intuitively, stronger bias from sender 2 makes sender 1's report relatively more credible, thereby expanding the range of states where the decision maker follows sender 1.

2.9.2 Comparative Statics on the Swing (Higher Costs)

Similarly, for changes in k_2:

$$\frac{ds(r_1)}{dk_2} = -\frac{\frac{dG_s}{dk_2}}{\frac{dG_s}{ds(r_1)}},$$

where $\frac{dG_s(r_1, s(r_1))}{dk_2} = \mathbb{1}\left\{\underline{r}_2^{-1}(s(r_1)) < r_1\right\} T(\underline{r}_2^{-1}(s(r_1)), \cdot) \frac{d\underline{r}_2^{-1}(s(r_1))}{dk_2}.$

Since

$$\frac{d\underline{r}_2^{-1}(r_2)}{dk_2} = \frac{\sqrt{1 + 4k_2(\tau_2 - r_2)} - [1 + 2k_2(\tau_2 - r_2)]}{2k_2^2\sqrt{1 + 4k_2(\tau_2 - r_2)}} < 0,$$

we obtain that

$$\frac{ds(r_1)}{dk_2} = \begin{cases} \geq 0 & \text{if } \underline{r}_2^{-1}(s(r_1)) < r_1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, when sender 2's misreporting becomes cheaper (lower k_2), the decision maker becomes more skeptical of high reports, and the swing report shifts downward for extreme messages.

2.9.3 Other Comparative Statics: Linear Costs

For completeness, consider linear misreporting costs $C_j(r_j, \theta) = |r_j - \theta|$. In this case, the derivative $\frac{dG_s}{ds(r_1)} = 0$, so the IFT cannot be applied directly. With uniform priors and $u_j(\theta) = \theta - \tau_j$, the swing condition becomes:

$$\int \frac{\theta}{(\theta - \tau_1)(\theta - \tau_2)}\, d\theta = 0,$$

which implies

$$\frac{\tau_1 \log\left(\frac{r-\tau_1}{s-\tau_1}\right) + \tau_2 \log\left(\frac{s-\tau_2}{r-\tau_2}\right)}{\tau_1 - \tau_2} = 0.$$

Using $\bar{r}_1(\theta) = 2\theta - \tau_1$ gives

$$\theta_l = \frac{2^{\tau_1/\tau_2} - \tau_1 - \tau_2}{2^{\tau_1/\tau_2} - 2}.$$

Thus, $d\theta_l/d\tau_2 > 0$ for all finite τ_2, but $\lim_{\tau_2 \to \infty} \theta_l = \tau_1(1 - \log 2) < 0$: even extreme bias does not yield full revelation when costs are linear.

2.9.4 Overall

Overall, these results imply that in Adversarial Equilibria:

$$\frac{d\theta_l}{d\tau_2} > 0, \quad \frac{d\theta_l}{dk_2} \geq 0,$$

and hence the receiver's welfare under competition weakly improves as additional senders become more biased or less costly. Competition thus acts as a force toward greater information revelation.

Chapter 3

The Shape of Moral Satisfaction

This chapter is based on Sanesi and Del Mastio, 2025. Minor editorial assistance, limited to text refinement, was performed using AI-based tools.

3.1 Introduction

Why do some individuals behave consistently across moral decisions, while others seem to relax after a good deed? Early research emphasized moral consistency: the "foot-in-the-door" effect (Freedman and Fraser, 1966), repeated helping (Kraut, 1973; Beaman et al., 1983; Burger, 1999), and theories of self-perception (Festinger, 1957; Gawronski and Strack, 2012) all suggested that prosocial behavior is self-reinforcing. More recent work, however, highlights moral licensing, whereby past moral acts reduce subsequent prosociality (Benoit Monin and Miller, 2001; Merritt, Effron, and Benoît Monin, 2010; Jordan, Mullen, and Murnighan, 2011; Brañas-Garza et al., 2013; Ploner and Regner, 2013; Sachdeva, Iliev, and Medin, 2009). Reviews document evidence for both patterns but emphasize that few studies examine them jointly (Mullen and Benoît Monin, 2016; Ferguson et al., 2024), and that theoretical frameworks reconciling these opposing dynamics remain scarce. To make some progress in this

direction, we begin by distinguishing the motives that drive moral behavior.

There are several reasons why people perform good deeds, which can be classified into two main categories. First, a good action is chosen because of the outcomes it produces, like increasing the payoffs of other individuals. Research on social preferences shows that people can be altruistic or inequity-averse, meaning their utility depends not only on their own payoffs but also on those of others (Fehr and Schmidt, 1999). Second, beyond outcome-based preferences, individuals may be motivated by internal rewards associated with behaving in ways that align with their values or self-concept. We refer to these internal rewards as *moral satisfaction* and we model it as an additional component of the utility function. The key feature of moral satisfaction is how its marginal return varies with prior moral behavior. If the marginal return to a moral act is increasing, engaging in moral behavior raises the incentive to behave morally again: each additional moral action strengthens future moral motivation, generating moral consistency. Conversely, if the marginal return is decreasing, performing a moral act reduces the incentive for subsequent moral behavior: past good deeds partially exhaust moral motivation, giving rise to moral licensing.

Our mechanism differs from the main explanations proposed in the literature on moral licensing and moral consistency. A first class of accounts explains these patterns through changes in self-image, self-signaling, or moral credentials: after a good deed, individuals update their beliefs about their own moral type or feel entitled to relax without threatening that image. A second class emphasizes consistency motives, such as self-perception, cognitive dissonance, or the desire to behave in line with a previously enacted identity. In both cases, past behavior matters because it changes beliefs, identity, or justification. By contrast, we abstract from belief updating and identity management, and focus instead on the shape of the utility directly generated by moral action itself. In our framework, the same reduced-form object, moral satisfaction, can generate either moral consistency or moral licensing depending on whether its marginal return is locally increasing or decreasing. This perspective

complements existing accounts by providing a unified decision-theoretic representation of both dynamics and by making the curvature of moral utility an empirically testable object.

This paper establishes, to the best of our knowledge, the first link between the shape of moral satisfaction, conceptualized as the hedonic utility derived from acting in accordance with one's personal normative beliefs (Bicchieri, 2016), and dynamic patterns of moral behavior. Our central research question is whether the moral satisfaction component of the utility function exhibits diminishing or increasing marginal returns to moral action, and whether these returns are heterogeneous across individuals. This question is crucial because the curvature of moral satisfaction directly determines whether past moral behavior crowds out or crowds in future moral effort, giving rise to moral licensing or moral consistency, respectively. As a secondary question, we examine whether moral satisfaction is domain-sensitive, that is, whether it depends on the type of moral action performed. By "type," we refer to the everyday-life domain to which each action pertains.

We develop a clean decision-theoretic framework and an empirical test to study the shape of moral satisfaction and its heterogeneity in the population. In an online experiment, we hold material incentives constant and vary only the moral framing. In the first phase, participants in two treatment groups are credited for a fixed donation made "thanks to their participation" and are asked to choose one of four non-consequential statements describing the meaning they attach to that donation. Each statement corresponds to a moral domain (e.g., environmental protection, social justice, or health promotion). Control participants are simply informed that the same donation has been made, without being credited for it. In the second phase, all participants receive the same endowment and decide how much to donate in a charity dictator game. We vary only the framing attached to the Phase-2 donation opportunity. Participants who selected a statement in Phase 1 are assigned either to a Same Domain condition, in which they keep the same statement in Phase 2, or to a Cross Domain condition, in which they select a different statement from the remaining options. Participants in the Con-

trol condition do not select a statement in Phase 1 and select one for the first time in Phase 2. Comparing Phase-2 donations across these conditions allows us to identify whether experiencing moral satisfaction, by being credited for an initial donation, affects the marginal moral utility of donating again, and whether moral satisfaction transfers across domains.

The nature of our research questions requires a between-subjects design. When individuals are aware that they are making multiple morally relevant choices, concerns about appearing consistent, both to themselves and to the experimenter, may mechanically induce coherence across decisions (Guadagno and Cialdini, 2010). Such behavior would confound genuine intertemporal moral dynamics with self-presentation and demand effects.

Identifying heterogeneity in the moral component of the utility function within a between-subjects design is not trivial. We address this challenge by stratifying participants based on the moral connotation of the language they use in open-ended responses, which serves as a proxy for their overall level of moral engagement. Language provides a natural behavioral trace of moral cognition: through repeated exposure to morally connoted experiences, individuals internalize moral norms that systematically shape their word choice and evaluative framing (Kennedy et al., 2021; Ramezani et al., 2024). By restricting comparisons to groups with the same level of overall moral engagement, we can isolate differences in the marginal moral utility of action across groups.

Our results regarding our main research question replicate the concavity observed among highly morally engaged participants in Carpenter (2021), while also revealing convexity among less morally engaged individuals. The curvature of moral satisfaction provides a unified explanation for these opposing patterns of dynamic moral behavior. Among low morally engaged participants, increasing marginal moral utility generates moral consistency (being credited with an initial moral act increases subsequent giving), whereas among highly morally engaged participants, diminishing marginal moral utility generates moral licensing after an initial moral act.

Our secondary question concerns the multifaceted nature of moral behavior. The experimental evidence suggests that moral satisfaction accumulates across domains, with individuals treating moral actions as substitutable sources of hedonic utility rather than as distinct and complementary moral goods.

Finally, we suggest interpreting these patterns as manifestations of a single underlying mechanism: an S-shaped function of moral satisfaction. At low levels of moral engagement, convexity generates moral consistency, whereas at higher levels, concavity gives rise to moral licensing. This framework not only reconciles opposing dynamic patterns within domains but also accounts for the fungibility of moral behavior across domains. When moral satisfaction is transferable, individuals treat moral acts in different domains as substitutes along the same underlying moral-utility curve. The heterogeneity observed in our experiment is therefore consistent with participants occupying different initial positions along the same non-monotonic moral satisfaction curve, rather than possessing fundamentally different moral utility functions. Individuals do not differ in nature, but in the moral experiences they bring into the decision context. Our language-based measure captures this variation as differences in baseline moral experience.

Taken together, these findings underscore the need for greater attention to the design and allocation of morally relevant incentives and nudges. Because the curvature of moral satisfaction varies across individuals[1], the same intervention may elicit opposite behavioral responses, strengthening moral engagement for some while inducing moral licensing in others. This heterogeneity implies that moral incentives are not universally effective: their impact depends on individuals' overall moral experience and on the broader context in which moral choices are presented.

Beyond individual behavior, our results highlight the strategic role of firms and institutions in shaping the moral choice set. By "selling" moral satisfaction, through campaigns that frame consumption or sym-

[1]Even if the moral satisfaction component of utility is common across individuals, its local curvature depends on the individual's accumulated moral experience.

bolic gestures as virtuous, firms effectively participate in a market for moral utility. When consumers treat moral acts as substitutable, these commercialized forms of moral satisfaction can crowd out welfare-enhancing behaviors, diverting limited moral motivation toward low-impact activities. In such cases, firms capture moral rents by offering hedonic rewards without delivering genuine social outcomes. From a welfare perspective, this raises a fundamental allocation problem. If moral motivation is a scarce resource, a social planner faces the challenge of deploying it toward actions with the highest social return, while firms may strategically appropriate it to create perceived virtue without real externalities. Phenomena such as greenwashing, social washing, and virtue signaling can thus be interpreted as equilibrium outcomes in markets where moral satisfaction itself is commodified. Recognizing moral satisfaction as both a private and positional good reframes traditional policy debates: interventions must not only enhance prosocial payoffs but also protect the moral incentive space from dilution by purely symbolic appeals.

The remainder of the paper is organized as follows. Section 3.2 introduces the theoretical framework. Section 3.3 formalizes the hypotheses regarding the shape of the moral satisfaction function, grounded in prior empirical and theoretical evidence. Section 3.4 details the design of the online experiment, while Section 3.5 presents the results and discusses their theoretical implications. Section 3.6 places our contribution in the relevant literature. Finally, Section 3.7 concludes and outlines directions for future research.

3.2 Decision-Theoretic Framework

This section introduces a decision-making model for an individual choosing among alternatives, building on the results of Alger and Weibull, 2013, the evidence brought by Bénabou, Falk, and Henkel, 2024 and Capraro, Halpern, and Perc, 2024, and the formalization of Minardi, Wang, and Gilboa, 2024. This model can also be interpreted in terms of the warm glow introduced by Andreoni, 1990, identity value presented by Akerlof and Kranton, 2000, or self-image by Bénabou and Tirole, 2006.

The alternatives are action bundles x in X, which is a closed and convex subset of $\mathbb{R}^n_+$. This means that the decision-maker (dm) chooses a bundle prescribing a level of intensity x_j for each action $j \in \{1, ..., n\}$.[2] Section 3.8.2 describes more clearly how this formalization extends to the case of a strategically interactive situation.

The dm has a hardwired personal norm indicating if action j is morally recommended under each one of the K ethical domains [3]. The personal normative beliefs across the multiple ethical domains of the dm are represented by the matrix $D_{n\times K}$.

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1k} & \dots & d_{1K} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{j1} & d_{j2} & \dots & d_{jk} & \dots & d_{jK} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nk} & \dots & d_{nK} \end{bmatrix}$$

[2] Here we talk about the intensity of actions. We believe that for each action, there is a quantitative dimension that can be chosen by the decision maker. For example, if the action is the consumption of a good, the quantity would represent the intensity. In the experiment, the actions in the bundle of the decision-makers are donations to a charity; there, the intensity of the action is the amount of money donated.

[3] It is important to clarify that by domains, we mean concrete, everyday situations where individuals face moral choices. Examples include environmentally responsible behaviors, helping less fortunate individuals, supporting social justice initiatives, assisting victims of natural disasters, or providing healthcare to elderly people. Each of these situations represents a specific context in which people make morally relevant decisions, guided by their personal normative beliefs about what actions are morally appropriate. For further clarifications, see Appendix 3.8.1.

For each moral domain $k \in \{1, ..., K\}$ (e.g., environment, human solidarity, social justice, etc.), the personal norm indicates which actions are considered morally valuable within that domain. Accordingly, for each action j, there is a binary indicator $d_{jk} \in \{0, 1\}$ that takes value 1 if the action j is deemed morally valuable in domain k, and 0 otherwise.[4] The indicator d_{jk} thus encodes the domain-specific normative belief about action j, while the vector d_k aggregates all the moral prescriptions under domain k. The full set of personal norms can be represented as a matrix D, which captures the moral value of each action across all domains. This matrix is assumed to be a hardwired characteristic of the decision maker: it reflects deep-rooted values shaped by past experiences, cultural background, and social influences, but is not directly chosen or manipulable by the decision maker. However, for an action that is aligned with the decision maker's personal norms to generate moral satisfaction–the intrinsic reward of doing the "right thing"–another element is required: framing. The presentation of available alternatives can highlight (or obscure) their moral dimension. Thus, for each action j we introduce a binary label $l_{jk} \in \{0, 1\}$, which acts as a spotlight: when $l_{jk} = 1$, the moral aspect of action j under domain k is made salient, enabling the moral satisfaction mechanism; when $l_{jk} = 0$, even if the action aligns with personal norms ($d_{jk} = 1$), it does not generate moral satisfaction because the moral relevance is not perceived.[5]

The full set of moral-saliency labels is captured by the matrix $L_{\{n \times K\}}$, where each element l_{jk} indicates whether the moral dimension of action j under domain k is activated through framing. These labels reflect how the decision context, especially the language and presentation of choices, shapes the moral perceptions of the decision maker without affecting

[4] A possible extension could allow $d_{jk} \in [-1, 1]$, to capture the strength of normative beliefs, including the possibility of immoral actions. However, for the purposes of this paper, we restrict d_{jk} to be a binary indicator, taking the value 1 if the action is morally valuable, and 0 otherwise, thereby excluding negative or intermediate values.

[5] An extension could allow for continuous values x, to capture varying degrees of moral salience in the framing. Another potential extension considers x as manipulable by external agents, such as a seller, who could strategically emphasize or downplay moral aspects to influence behavior.

monetary outcomes.

$$L = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1k} & \dots & l_{1K} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{j1} & l_{j2} & \dots & l_{jk} & \dots & l_{jK} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \dots & l_{nk} & \dots & l_{nK} \end{bmatrix}$$

Summarizing, we consider the following components of the model:

- The **moral saliency of the framing** with which each action is presented to the decision maker, described by the matrix L, where each element l_{jk} reflects whether the moral dimension of action j is made salient through framing.
- A fixed set of domain-specific **personal norms**, represented by the matrix $D \in \{0,1\}^{n \times K}$, where each element d_{jk} indicates whether good j is morally valuable in domain k.
- An **action bundle** $x \in \mathbb{R}^n$, chosen by the decision maker from the feasible set of alternatives $X \subset \mathbb{R}^n$.

These elements jointly determine how moral considerations, when made salient through framing, influence the decision maker's choices beyond standard economic preferences.

To understand how these elements enter the utility function of the decision-maker, let us consider the case $K = 1$, which means that there is only one moral domain. If the vectors d and l are known and constant, it is possible to partition X, the set of available bundles, into two subsets:

$$X^0 := \left\{ x \in X \;\middle|\; \sum_{j=1}^{n} x_j \cdot l_j \cdot d_j = 0 \right\},$$

$$X^1 := \left\{ x \in X \;\middle|\; \sum_{j=1}^{n} x_j \cdot l_j \cdot d_j > 0 \right\}.$$

In this way, X^1 represents the set of alternatives that, if chosen under framing l, give the decision-maker some moral satisfaction under domain k. This moral satisfaction should be understood as an individual deriving fulfillment from committing a moral action simply by following a given norm, regardless of the consequences, whether good or bad.[6]

The idea is the following. Each action can either be framed as morally salient or not, and can be considered morally valuable or not according to the decision-maker's personal normative beliefs. A decision-maker experiences moral satisfaction if she chooses a bundle that assigns positive intensity to at least one action that is both framed as morally salient and considered morally valuable under her personal norms.

The intensity with which the alternative is consumed does not alter the fact that she is perceiving the moral premium, but may indeed affect its magnitude. [7]

Things become more interesting when considering more than one moral domain at work on the same set of alternatives ($K > 1$). Individuals derive satisfaction from choosing morally connoted alternatives that they perceive as positive. Therefore, the moral premium is activated whenever a morally framed action, considered morally valuable under at least one domain, is performed.[8] Formally, let X^1 be the set of bundles that generate moral satisfaction. In the case of multiple moral domains, X^1 includes all alternatives where at least one good is both morally framed and considered morally valuable under at least one

[6]In this paper, we focus on the sense of satisfaction coming from fulfilling moral principles. We exclude from this analysis the sense of guilt deriving from the explicit violation of a moral norm. This aligns with the experimental design, where participants are asked to perform only positively connoted acts, such as donating money to a charity.

[7]When $d_{jk} = 1$ but $l_{jk} = 0$, it means that the consumption of that good should be considered morally valuable, but the lack of moral connotation of the framing nullifies the moral satisfaction of the dm. If $d_{jk} = 0$ but $l_{jk} = 1$, it means that the framework is morally connoted, but the dm does not believe that it is a moral duty to consume good j.

[8]We assume that the framing and the personal normative belief must be in the same domain in order to produce the satisfaction. This means that if an action is morally framed under one domain but considered good in another, it will not produce satisfaction when performed.

moral domain, and is consumed in positive quantity:

$$X^1 := \left\{ x \in X \;\middle|\; \sum_{k=1}^{K} \sum_{j=1}^{n} x_j \cdot l_{jk} \cdot d_{jk} > 0 \right\}.$$

The total utility of a decision-maker choosing bundle x can then be formalized as follows:

$$U_i(x) = \begin{cases} \underbrace{u_i(x)}_{\text{selfish and other regarding outcome based utility}} + \underbrace{\gamma(\sum_{k=1}^{K} \sum_{j=1}^{n} x_j \cdot l_{jk} \cdot d_{jk})}_{\text{moral premium}} & \text{if } x \in X^1 \\ \underbrace{u_i(x)}_{\text{selfish and other regarding outcome based utility}} & \text{otherwise} \end{cases}$$

Where the function $\gamma(\cdot)$ represents the *moral premium*. Consistent with the theoretical and empirical literature reviewed in Section 3.1, we assume that $\gamma(\cdot)$ is an increasing function. The argument of γ corresponds to the sum of the intensity of actions framed in a moral way and perceived by the individual as appropriate. In Section 3.3, we propose specific hypotheses regarding the second derivatives of these functions, which yield testable predictions for the experiment presented in Section 3.4.

3.3 Hypotheses on the shape of the moral premium

The descriptive model just introduced in section 3.2 allows us to formalize a set of hypotheses regarding the shape of the moral premium, grounded in findings from psychology and behavioral economics [9]

Formally, the utility function is defined as:

$$U_i(x) = \begin{cases} u_i(x) + \gamma(a_1, a_2, \ldots, a_K) & \text{if } x \in X^1, \\ u_i(x) & \text{otherwise,} \end{cases}$$

where $a_k = \sum_{j=1}^{n} x_j \, l_{jk} \, d_{jk}$ represents the total moral action in domain k.

Our hypotheses will specify something regarding the second derivative of function $\gamma(\cdot)$.

(I) – *Concavity*: Decreasing marginal returns from moral actions The first hypothesis for the shape of $\gamma(\cdot)$ is that it is a concave function. Formally,

$$\frac{\partial^2 \gamma}{\partial a^2} < 0.$$

The behavioral interpretation of this hypothesis is that the satisfaction derived from doing what one believes to be the right thing, when such an action is morally framed, increases with the intensity and number of moral actions performed, albeit with decreasing marginal returns. In a dynamic context, if this hypothesis holds, the model predicts that individuals' willingness to perform further moral actions diminishes as they accumulate more such actions. This reduction in moral motivation is consistent with the phenomenon of *moral licensing* (Merritt, Effron, and Benoît Monin, 2010). The following examples should clarify the behavioral consequences of such a hypothesis.

[9] The hypotheses were preregistered at the Open Science Framework (OSF): `https://osf.io/9hekj`.

Example 1. A person donates 5 euros to a charity and feels a sense of moral satisfaction. Donating an additional 5 euros still increases that satisfaction, but not as much as the first donation did. After a certain point, further donations might feel redundant or less emotionally rewarding.

Example 2. Someone reduces their meat consumption for environmental reasons. The first few meatless meals provide a strong moral signal and emotional reward. As the behavior becomes routine, the additional moral satisfaction from each subsequent meatless meal diminishes.

(II) – *Convexity*: Increasing marginal returns from moral actions The second hypothesis regarding the shape of $\gamma(\cdot)$ is that it is a convex function. Formally,

$$\frac{\partial^2 \gamma}{\partial a^2} > 0.$$

The behavioral interpretation of this hypothesis is that the more moral actions an individual undertakes, the greater is the marginal contribution of additional actions to moral satisfaction. In this case, once the individual performs a moral act, she experiences an increased incentive to behave morally again. This behavioral pattern is consistent with the concept of *consistency* in moral behavior (Kraut, 1973).

Example 1. A person who donates a small amount to a charity may feel some moral reward, but this satisfaction could grow disproportionately if they continue donating or increase the amount over time. The more they give, the stronger their identity as a "generous" or "moral" person becomes, reinforcing future generosity.

Example 2. Someone who begins volunteering for an environmental cause might initially derive modest satisfaction from participating. However, as they become more involved, taking leadership roles, organizing events, or changing personal habits, the moral fulfillment deepens, encouraging sustained and even escalating commitment.

(III) – *Linearity*: Constant marginal returns from moral actions For completeness, the third hypothesis on the shape of $\gamma(\cdot)$ is that it is a linear

function. Formally,

$$\frac{\partial^2 \gamma}{\partial a^2} = 0.$$

If this is the case, we should find no change in the marginal return from moral actions in a dynamic context. If $\gamma(\cdot)$ is linear, it should play no role in the mixed evidence observed in the literature on dynamic moral behavior presented in Section 3.1, suggesting these phenomena should be investigated from an outcome-based point of view rather than one based on internal moral dynamics.

The following examples illustrate this behavioral pattern without assuming that individuals are indifferent to cumulative or distributive consequences.

Example 1. A person donates a fixed amount to charity every month without changing behavior elsewhere. Each act is treated as morally neutral, with no spillover in motivation from past or future donations.

Example 2. A consumer consistently buys fair trade products but does not adjust their purchasing decisions based on prior consumption. Moral behavior follows a constant rule, uninfluenced by past behavior.

(IV) – *Domain Fungibility*: insensitivity to domain specification In the presence of multiple moral domains, distinct areas of life in which individuals can act morally, the fourth hypothesis of *domain insensitivity* holds that the moral satisfaction component of utility, represented by $\gamma(\cdot)$, depends on the total amount of morally connoted and normatively aligned actions, regardless of their distribution across domains. Formally, for domains $k \neq h$ and actions a_k, a_h:

$$\frac{\partial^2 \gamma}{\partial a_k \, \partial a_h} = \frac{\partial^2 \gamma}{\partial a_k^2}, \quad \forall\, k \neq h.$$

This implies that moral actions in different domains neither reinforce nor offset one another: $\gamma(\cdot)$ exhibits no complementarity or substitution across domains. Domain insensitivity concerns only the moral satisfaction derived from acting according to one's own moral norms, and not the outcome-based component of utility, $u_i(x)$, which may be affected

by domain specification. For instance, an individual might value welfare improvements in an environmental context than in an environmental one, a difference captured by $u_i(x)$ rather than by $\gamma(\cdot)$.

Example 1. An individual with a convex and domain-insensitive $\gamma(\cdot)$ may experience an increased moral incentive to recycle diligently after attending church. This person derives the same type of moral satisfaction from both actions, as each aligns with their moral beliefs.

Example 2. An individual with a concave and domain-insensitive $\gamma(\cdot)$ may experience a reduced moral incentive to donate to a child-care charity after committing not to eat meat for environmental reasons. This satisfaction stems from acting in accordance with their moral norms, regardless of whether they consider one domain more impactful than another in terms of outcomes.

Domain insensitivity thus implies a kind of psychological fungibility across domains with respect to norm-based fulfillment, even if the agent's outcome-based preferences are domain-specific.

(V) – *Domain-Specific Utility*: Preference for Same-Domain The fifth hypothesis assumes that individuals experience stronger marginal returns to moral actions within a single domain than across different domains. While our focus is on the positive domain of moral utility, this idea is consistent with Fanghella and Thøgersen, 2022, who show that recalled immoral actions primarily trigger moral cleansing within the same domain, with individuals engaging more in compensatory behavior in the domain of the original transgression rather than switching to another domain. Moral satisfaction derived from actions in one domain may slightly affect incentives in others, but these cross-domain effects are weaker. This structure produces behavioral consistency within domains and a tendency toward specialization in one preferred moral area.

The key assumption characterizing this hypothesis is that moral reinforcement occurs mainly within domains:

$$\frac{\partial^2 \gamma}{\partial a_k \, \partial a_h} < \frac{\partial^2 \gamma}{\partial a_k^2}, \quad \forall \, k \neq h.$$

This implies that while a moral action in one domain can slightly increase moral motivation in others, the marginal effect of same-domain actions remains dominant.

This specification emphasizes *intra-domain coherence*: agents derive greater *norm-based* moral satisfaction from focusing their morally connoted actions within a specific domain rather than distributing effort evenly across multiple ones. Importantly, this assumption pertains exclusively to the *normative component* of utility, i.e., the moral satisfaction captured by $\gamma(\cdot)$.

Example 1. An individual may repeatedly engage in environmental behaviors (such as biking to work, avoiding plastic, or composting) deriving increasing norm-based moral satisfaction from focusing effort within that domain. They may still donate to charity seeking moral satisfaction, but this decision is only weakly affected by their prior environmental actions.

Example 2. An individual may prefer to focus all their resources on supporting civil rights causes (through protests, petitions, or donations) rather than also contributing to cancer research. This is not because they do not care about cancer research, but because the moral returns are maximized when effort is concentrated in a single domain rather than spread across multiple domains.

(VI) – *Domain Specific Utility*: Preference for Cross-Domain Lastly, the sixth hypothesis proposes that the moral premium $\gamma(\cdot)$ captures a form of complementarity across moral domains. In this setting, $\gamma(\cdot)$ is an increasing function of all moral actions, and its cross-partial derivatives with respect to actions in different domains are not only positive but also greater than the corresponding second derivatives within each domain. Formally, for domains $k \neq h$ and actions a_k, a_h:

$$\frac{\partial^2 \gamma}{\partial a_k \, \partial a_h} > \frac{\partial^2 \gamma}{\partial a_k^2}, \quad \forall \, k \neq h.$$

This condition implies that a good deed in one domain increases the moral incentive to behave well in another domain more strongly than it

reinforces further behavior within the same domain. Individuals, therefore, derive greater moral satisfaction from maintaining a balanced and diversified moral profile across different areas of life. This is in line with Peetz and Howard, 2022, which shows that people prefer to diversify their prosocial efforts across different types of help. Positive and relatively larger cross-partial derivatives indicate a strong degree of complementarity between domains.

Here, $\gamma(\cdot)$ is an increasing function such that cross-domain interactions are stronger than intra-domain effects.

Example 1. An individual may feel more motivated to recycle after attending church than after biking to work. The moral action in the religious domain reinforces moral motivation in the environmental domain more strongly than within-domain consistency alone would predict.

Example 2. A person may derive norm-based satisfaction from maintaining a "balanced moral portfolio", for instance, buying fair-trade products, volunteering at a shelter, and using public transport. If they stop engaging in one of these domains, their overall moral satisfaction $\gamma(\cdot)$ declines, even if they intensify efforts in other domains, because cross-domain complementarities amplify the contribution of diversification.

In the next section, we will test the validity of the listed hypotheses. We emphasize that these considerations pertain to the norm-based component of the utility function. Therefore, while any changes in the outcome-based sphere may trigger consistency or compensation effects, we aim to isolate the impact on the norm-based dimension by holding the outcome-based factors constant.

3.4 Experiment

To study how the shape of the norm-based component of the utility function varies with adherence to different morally connoted domains, our experimental design holds outcome-based motives constant while exogenously manipulating norm-based ones. Specifically, we vary the moral meaning attached to an action without changing its material consequences. This approach follows earlier work that isolates the intrinsic moral value of actions, such as donating, independently of their outcomes, including Burum, Nowak, and Hoffman (2020) and Crumpler and P. J. Grossman (2008).

3.4.1 Design

We design an online experiment in which participants are randomly assigned to one of three groups: Control (C), Same-Domain (SD), or Cross-Domain (CD), as summarized in Table 2. Participants complete two sequential phases. Because Phase 2 follows Phase 1 in quick succession, the moral frame made salient in Phase 1 remains available when participants make the Phase 2 decision. Before the experimental tasks begin, all participants read the same detailed description of the activities of the Italian Red Cross (Appendix 3.9.2), so that information provision does not vary across treatments. Participants are also informed that, at the end of the study, one out of every 100 participants is randomly selected and that this participant's decisions are implemented for payment. All remaining participants receive a fixed participation fee of 0.50 € (Appendix 3.9.1).

First phase. In Phase 1, participants in the **Control group** are informed that the experimenter will donate 100 € to the Italian Red Cross.[10] Participants are told that they will not observe the precise use of the funds and that the Italian Red Cross will allocate the donation across ongoing projects in different sectors, in line with its standard budget policy (Appendix 3.9.2).

[10] We use the term *donation* to refer to any transfer of money to the Italian Red Cross, whether made by the experimenter or by the participant.

Table 2: Experimental design with two phases and three treatments.

Control Group (n=199)	**CD Group (n=200)**	**SD Group (n=211)**
Informed of 100 € donation	Entitled of 100 € donation + choose a statement THANKS POPUP	Entitled of 100 € donation + choose a statement THANKS POPUP
$u(0,100) + \gamma(prior)$	$u(0,100) + \gamma(prior + 100_k)$	$u(0,100) + \gamma(prior + 100_k)$
Charity DG + choose a statement	Charity DG + choose **different** statement	Charity DG + **same** statement
$Eu(100 - x^C, 100 + x^C) + \gamma(prior + x_k^C)$	$Eu(100 - x^{CD}, 100 + x^{CD}) + \gamma(prior + 100_k + x_{\not k}^{CD})$	$Eu(100 - x^{SD}, 100 + x^{SD}) + \gamma(prior + 100_k + x_k^{SD})$
Questionnaire and open text		

Notes: Utility evaluation at each stage of the experiment is provided. Outcome-based utility u is expected due to random payment and depends on both the participant's payoff and the Italian Red Cross' payoff. The moral premium γ depends on the – unobservable to the experimenter – moral actions performed by the participant prior to entering the experiment ($prior$) and the morally framed actions taken within the experiment – donations with the statement choice. The variable k denotes the statement chosen by the participant.
The Control Group faces just one morally connoted choice, the CD Group faces two morally connoted choices in two different domains, and the SD Group faces two morally connoted choices in the same domain.

Participants in the **Same-Domain** and **Cross-Domain** groups are instead informed that the same donation of 100 € will be made *thanks to their participation in the experiment*.[11] As in the Control group, we explicitly state that the allocation of funds is entirely determined by the charity and is unaffected by participants' choices.

Participants in the SD and CD groups then select a statement to communicate the meaning they attach to the donation. This choice has no outcome-based consequences and does not affect how the charity uses the funds. The design builds on the *Statement Choice* experiment in Bénabou, Falk, and Henkel (2024), which shows that individuals attach intrinsic value to the moral meaning conveyed by a statement even when it does

[11] To evoke a sense of moral ownership, we use the following wording: *"Thanks to your participation in this experiment, a donation of 100 € will be made to the Italian Red Cross. We ask you to select a statement to authorize the donation and tell us the meaning you attach to it. Your choice will remain confidential and will only be used for this research."*

not affect outcomes.[12] We reiterate that statement selection has no effect on the allocation of the donation and is collected solely for research purposes. Participants also complete a comprehension check before entering the experimental tasks to confirm this understanding (Appendix 3.9.3).

Participants choose one of the following four statements to authorize the donation and indicate the meaning they wish to assign to it:

- By donating, I want to support families whose lives have been disrupted by **migration**, ensuring they have access to essential services.

- By donating, I want to support people when natural disasters strike, providing help before, during, and after **emergencies**.

- By donating, I want to support initiatives that protect the physical and mental **health** of all, with particular attention to the elderly and people with disabilities.

- By donating, I want to support people struggling with **poverty** and the rising cost of living by offering both practical assistance and emotional support.

Each statement is accompanied by an image to enhance moral salience.

After the selection, participants receive a personalized pop-up message:

> *"Thank you! Now we know that [selected domain] is important to you."*

The image associated with the chosen statement is displayed again in the pop-up.

[12] In Bénabou, Falk, and Henkel (2024), participants were willing to reduce their donation to select "I support the preservation and protection of the environment" rather than "I support the destruction of the environment," suggesting that statement choice carries perceived moral value. In our interface, each statement is paired with an image to increase salience.

Second phase. In Phase 2, all participants receive an endowment of 100€ and decide how much to keep and how much to donate to the Italian Red Cross. The donation decision is framed by a statement that depends on the assigned group:

- **Control:** participants select one of the four statements (for the first time).
- **Same-Domain:** participants are shown again the same statement selected in Phase 1.
- **Cross-Domain:** participants select one of the three statements not chosen in Phase 1.

After the statement is selected or displayed, participants use a slider to allocate the 100€ between themselves and the Italian Red Cross.

Rationale. The monetary environment is the same across treatments. In Phase 1, the Italian Red Cross receives the same amount in all conditions. In Phase 2, all participants face the same feasible set of allocations between themselves and the charity. What differs across conditions is only the moral framing attached to donating: in Phase 1, SD and CD participants are credited for the donation and select a statement that fixes its moral meaning; in Phase 2, the donation opportunity is framed either in the same moral domain as before (SD), in a different domain (CD), or in a newly chosen domain (C). We interpret systematic differences in Phase 2 donations across conditions as responses to these norm-based manipulations, rather than to differences in material incentives or in the objective consequences of giving.

3.4.2 Questionnaire

After completing the main task, we measure some important variables as controls.

Personal and Social Normative Belief Elicitation We measure personal normative beliefs, that is, what individuals think they *should* do[13]. This allows us to control for participants' underlying inclinations toward the five moral domains explored in the previous section. To illustrate, we follow the approach of Bicchieri (2016), asking participants to indicate their level of agreement with statements such as:

> "Achieving fair access to health services for all is a goal that I feel it is my duty to help pursue."

Responses are collected using a 4-point Likert scale (ranging from *Strongly Disagree* to *Strongly Agree*), enabling us to capture subtle variations in participants' normative convictions. To mitigate order effects, the questions will be presented in a randomized order.

Further, we measure the social normative expectations of our participants (Bašić and Verrina, 2024). Given the important role that social norms play in shaping personal values, we consider it essential to also control for participants' perceptions of others' normative beliefs—that is, what they believe others think one should do.

Support for Italian Red Cross and other volunteer organizations We also measure participants' opinions on the activities of volunteer organizations in general, and the Italian Red Cross in particular. Participants are asked to rate their level of support on a scale from 1 to 10, where:

- 1 = "Do not support the work of voluntary organizations (Italian Red Cross) at all"
- 10 = "Totally support the work of voluntary organizations (Italian Red Cross)"

This measure allows us to control for participants' prior attitudes toward volunteer organizations and the Red Cross, but it also represents a proxy of prosociality.

[13]From a theoretical point of view, these questions ensure that d_{jk} takes the value 1 for the action donating to the Italian Red Cross, under the described domains.

3.4.3 Open text at the end of the experiment

Our experimental design follows a between-subject structure to avoid anchoring and consistency concerns that arise when individuals face repeated moral choices (Guadagno and Cialdini, 2010). Combined with randomized assignment to treatments, this allows us to rule out potential biases arising from heterogeneity in social preferences.

Nonetheless, the shape of the moral satisfaction function $\gamma(\cdot)$ may itself be heterogeneous across participants. This heterogeneity can arise through two distinct channels. First, $\gamma(\cdot)$ may be individual-specific, reflecting differences in baseline moral sensitivity or responsiveness to morally framed actions. Meaningful between-subject comparisons then require grouping participants with similar moral attitudes.

Alternatively, $\gamma(\cdot)$ may be common across individuals but exhibit a non-linear, potentially non-monotonic shape. Under this interpretation, the marginal moral satisfaction generated by a given action depends on an individual's position along the moral satisfaction domain. In particular, the curvature of $\gamma(\cdot)$ may differ across regions, implying that identical moral stimuli can generate increasing or diminishing marginal responses depending on prior moral engagement. We explore this second interpretation in greater detail in Section 3.5.1.

To empirically assess these possibilities, we elicit participants' motivations through an open-ended text question at the end of the experiment, asking them to explain the reasoning behind their decisions in their own words. We analyze these responses using text-based measures of moral content and use them to classify participants according to the moral connotation of their language. We interpret this classification as a proxy for participants' baseline level of moral engagement. Using this proxy, we group individuals into low- and high-moral-engagement categories, which allow us to infer features of the shape of $\gamma(\cdot)$ for each group.

The English translation of the corresponding survey prompt is reported below:

"Now we ask you to briefly explain the main reason behind

your choices in the three sections of the experiment. If you have any additional comments or reflections, feel free to share them here."

3.5 Results

The experiment allows us to test the hypotheses presented in Section 3.3. By comparing donations in Phase 2 across treatments, we can evaluate the curvature of moral satisfaction $\gamma(\cdot)$ and domain sensitivity. Table 3 summarizes the main patterns and their interpretations.

Table 3: Summary of hypotheses and predictions

Compatible Behavior	**Hypothesis**	**Formalization**
$x^C > \max\{x^{SD}, x^{CD}\}$	Concavity (I)	$\frac{\partial^2 \gamma}{\partial a^2} < 0$
$x^C < \max\{x^{SD}, x^{CD}\}$	Convexity (II)	$\frac{\partial^2 \gamma}{\partial a^2} > 0$
$x^{SD} = x^{CD} = x^C$	Linearity (III)	$\frac{\partial^2 \gamma}{\partial a^2} = 0$
$x^{SD} = x^{CD}$	Domain Fungibility (IV)	$\frac{\partial^2 \gamma}{\partial a_k \, \partial a_h} = \frac{\partial^2 \gamma}{\partial a_k^2}$
$x^{SD} > x^{CD}$	Preference for Same-Domain (V)	$\frac{\partial^2 \gamma}{\partial a_k \, \partial a_h} < \frac{\partial^2 \gamma}{\partial a_k^2}$
$x^{SD} < x^{CD}$	Preference for Cross-Domain (VI)	$\frac{\partial^2 \gamma}{\partial a_k \, \partial a_h} > \frac{\partial^2 \gamma}{\partial a_k^2}$

Notes: x^C, x^{CD}, x^{SD} represent donations in Phase 2 of the experiment under Control, Cross Domain and Same Domain conditions respectively.

The initially preregistered comparison reveals no significant difference in average donations across treatments.[14] This outcome may reflect linearity and domain fungibility, but it might also indicate that the moral satisfaction function $\gamma(\cdot)$ we are measuring varies across individuals rather than being homogeneous. To account for this potential heterogeneity, we adopt the strategy described in Section 3.5.1, clustering participants according to their motivational profiles. This analysis was not preregistered.

Our identification strategy proceeds in three steps. First, we compare donations in the Control condition with donations in the treatment conditions, Same-Domain and Cross-Domain, to identify the curvature of $\gamma(\cdot)$. If exposure to an additional moral stimulus increases subsequent

[14] The hypotheses and analysis plan were preregistered at the Open Science Framework: `https://osf.io/9hekj`.

donations, moral satisfaction exhibits increasing marginal returns, consistent with convexity. If it reduces subsequent donations, moral satisfaction exhibits decreasing marginal returns, consistent with concavity. No difference is consistent with linearity. Second, we compare behavior across the Same-Domain and Cross-Domain treatments to test whether moral satisfaction is domain-specific or fungible. Equal donations across these conditions indicate that moral actions are interchangeable across domains, whereas systematic differences reveal domain sensitivity. Third, to account for heterogeneity in responses, we exploit variation in participants' moral engagement, as captured by their language, and estimate treatment effects within more homogeneous subgroups. This approach allows us to isolate differences in the shape of $\gamma(\cdot)$ while holding constant underlying moral predispositions.

We structure the results in three parts. First, we address the issue of potential heterogeneity or non-monotonicity in $\gamma(\cdot)$. Second, we examine the curvature of $\gamma(\cdot)$ by comparing the Control and Cross-domain treatments. Third, we analyze domain sensitivity across different moral domains. A final discussion on the results closes this Section before the conclusions of the paper.

3.5.1 Dealing with heterogeneity of $\gamma(\cdot)$

Our design is between-subject, which prevents anchoring effects. Combined with randomized assignment to treatments, this allows us to rule out confounding influences arising from heterogeneity in social preferences. Nonetheless, as discussed in Section 3.3, the shape of $\gamma(\cdot)$ may vary across participants, particularly in their response to marginal increases.

Following Carpenter (2021), who analyzed curvature only among high–warm-glow individuals, we separate participants into two groups: those exhibiting high overall moral engagement and the remainder. To measure engagement, we elicit participants' motivations through open-ended, plain-text explanations of their decisions. Using text analysis, we classify these responses according to their moral connotation, which serves as a proxy

for moral engagement. This procedure allows us to cluster participants into low- and high-moral-satisfaction groups and to infer the shape of $\gamma(\cdot)$ for each group.

Language provides a natural behavioral trace of moral cognition. Moral psychology and computational linguistics show that individuals internalize moral norms that systematically shape their word choice and evaluative framing (Kennedy et al., 2021; Ramezani et al., 2024). Linguistic markers are trait-like and persist across contexts (Pennebaker, 2011), suggesting that the moral connotation of participants' explanations reflects their overall moral experience. Participants who spontaneously describe their actions in explicitly moral terms reveal a chronically activated moral self, whereas those using neutral or instrumental language signal lower moral engagement. This language-based classification thus provides an indirect but robust way to distinguish between high- and low-moral types, extending the approach of Carpenter (2021) beyond the immediate experimental context.

Given this, we cluster our analysis among participants using the same type of language. This ensures that, even in a between-subject design, comparisons are effectively restricted to individuals with similar underlying moral experience, isolating differences in $\gamma(\cdot)$ while controlling for variation in moral engagement.

We classify participants' written motivations (Section 3.4.3) as *morally connoted* or *non-morally connoted* using LLAMA (Touvron et al., 2023), a large language model prompted to detect explicit references to moral values, ethical principles, or duty.[15]

For illustration, the motivation "I chose what I believe is morally right to do" is classified as moral (LLAMA Moral Index = 1), whereas "All initiatives are commendable; the ones I selected seem more useful and effective" is classified as non-moral (LLAMA Moral Index = 0).

The classification using LLAMA to identify the two levels of prior moral engagement does not differ significantly across treatments (Con-

[15] As a robustness check, we also implemented a keyword-based classification relying on terms such as *justice*, *solidarity*, and *responsibility*, as reported in Appendix 3.10.2. Both methods produced qualitatively consistent results, although statistical significance was weaker for the keyword-based approach, consistent with its lower precision.

trol, Same Domain, and Cross Domain), as shown in Table 4 and confirmed by a chi-squared test ($\chi^2 = 0.6776$, $p = 0.4104$, $df = 1$).

Table 4: Contingency Table for Llama Moral Index by Treatment

Llama Moral Index	Control	Same Domain	Cross Domain
0	107	123	91
1	92	88	111

3.5.2 Curvature (HP I-III)

To verify the first three hypotheses concerning the curvature of the moral satisfaction function $\gamma(\cdot)$, we compare donations between the Control group and the experimental treatments.

An initial comparison reveals no significant difference in average donations across treatment groups, as shown in Table 5.

Table 5: Mann–Whitney U Test Results (Full Sample)

Comparison	U Statistic	p-value	Significance
Control vs. Cross Domain	19568.00	0.7709	n.s.
Control vs. Same Domain	19552.50	0.2227	n.s.
Cross Domain vs. Same Domain	19844.50	0.2905	n.s.

Notes: Significance levels are denoted as * $p < 0.05$, ** $p < 0.01$, n.s. = not significant.

This outcome may reflect linearity and domain fungibility, but it might also suggest that the moral satisfaction function $\gamma(\cdot)$ we are measuring varies across individuals rather than being homogeneous. To account for this potential heterogeneity, we adopt the strategy described in Section 3.5.1, clustering participants according to their motivational profiles.

Restricting the comparisons of donations in Phase 2 within the same language group, see Table 6, we find heterogeneous treatment effects between people using a morally connoted language and others. Among participants with a low level of moral connotation, donations are signif-

icantly higher in the treatment groups compared to the control. Conversely, among those with a high level of moral connotation, donations are lower in the treatment groups.

Table 6: Average Donations by Llama Moral Index and Treatment Group

	Treatment Group	Average Donation	Standard Error
Low Llama Index	Control	64.09	3.12
	Cross Domain	75.20	2.93
	Same Domain	75.26	2.70
High Llama Index	Control	74.16	3.12
	Cross Domain	66.51	2.84
	Same Domain	70.55	3.05

Table 7 reports the results of the Mann–Whitney U test, a non-parametric rank-based test that assesses whether values in one group tend to be larger than those in another, chosen here because the donation data are not normally distributed.

Table 7: Mann–Whitney U Test Results by Llama Moral Index Group

Comparison	U Statistic	p-value	Significance
Low Llama Moral Index Group			
Control vs. Cross Domain	3883.50	0.0131	**
Control vs. Same Domain	5306.50	0.0104	**
High Llama Moral Index Group			
Control vs. Cross Domain	5949.50	0.0406	*
Control vs. Same Domain	4417.00	0.2839	n.s.

Notes: Significance levels are denoted as * $p < 0.05$, ** $p < 0.01$, n.s. = not significant. The results for the Low Llama Moral Index group remain robust after applying the Bonferroni correction for multiple hypothesis testing (Bonferroni, 1936), with an adjusted significance level of $\hat{\alpha} = 0.025$.

Figure 9 illustrates the distributional shifts across treatments. For participants with a low moral index, both treatments first-order stochastically dominate the control group, indicating higher donation levels throughout the distribution. In contrast, the opposite pattern emerges among high–moral-index participants, where the control distribution dominates. The only exception is the Same-Domain condition within the high-moral group, where cumulative distributions partially overlap, consistent with the absence of significant differences reported earlier.

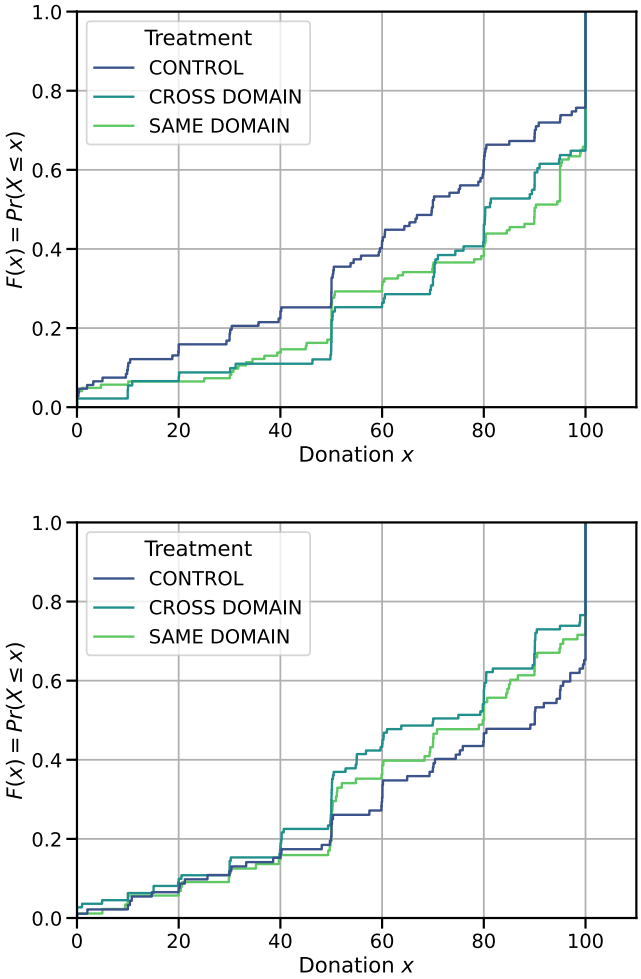


Figure 9: *CDFs of donations by morality index and treatment. Top: low morality; Bottom: high morality.*

Notes: CDFs show the proportion donating at most x euros.

A Tobit regression, controlling for the elicited support for IRC (Table 8), confirms the opposite effect. While the treatments reduce dona-

tions on average ($\beta = -11.50$, $p = 0.033$), their interaction with low morality is positive and significant ($\beta = 19.88$, $p = 0.010$). Translating coefficients into direct treatment effects, we see that high morality participants donate significantly less in the treatment (-11.5, $p = 0.033$), while low morality participants donate more ($+8.38$), though not significantly ($p = 0.131$).

Table 8: Tobit Estimates of Donation Behavior

	Coef.	SE	p-value
Treatment	-11.50	5.39	0.033*
Support for IRC	7.77	1.14	0.000**
Low morality	-7.30	5.51	0.185
Low morality × Treatment	19.88	7.74	0.010*
Low morality	8.38	5.56	0.131
High morality	-11.50	5.39	0.033*

Dependent variable: donation (0–100). Treatment is a dummy variable that takes value 1 if the subject is either in the Same Domain or Cross Domain group. Tobit model with censoring at 0 and 100.
*p<0.05; **p<0.01.

3.5.3 Ethical domains (HP IV-VI)

We examine domain effects by comparing Cross and Same domain conditions. Referring again to Tables 5 and 6, we can see that no statistical difference exists between donations in Same Domain and Cross Domain conditions, also when controlling for the level of LLama Moral Index as reported in Table 9.

Table 9: Mann–Whitney U Test Results by Llama Moral Index Group

Llama Moral Index	Comparison	U Statistic	p-value	Significance
Low	Cross vs. Same Domain	5529.00	0.8786	n.s.
High	Cross vs. Same Domain	4502.00	0.3403	n.s.

Notes: Significance levels are denoted as * $p < 0.05$, ** $p < 0.01$, n.s. = not significant.

Treatment effects are robust across domain specification. Differences between same- and cross-domain are not significant, indicating *domain fungibility*: participants respond primarily to moral action regardless of its domain.

3.5.4 Discussion

In this section, we discuss the results presented in Section 3.5 to provide a more comprehensive interpretation of the phenomenon. Table 10 summarizes the experimental evidence just presented.

Table 10: Moral Index, $\gamma(\cdot)$ Shape, and Interpretation

Moral Index (Donation)	$\gamma(\cdot)$ Shape	Interpretation
Low (Control $<$ Treatments)	Convex	Moral consistency
High (Control $>$ Treatments)	Concave	Self-licensing
(Cross Domain $\sim$ Same Domain)	Domain insensitivity	Cross-domain reinforcement

Notes: "Low" and "High" refer to clusters based on the moral index. Convexity implies increasing marginal returns (moral consistency), while concavity implies decreasing marginal returns (self-licensing). Domain insensitivity indicates participants respond to overall moral satisfaction rather than specific domains.

Section 3.5.2 shows that we identified two groups of individuals exhibiting markedly different behaviors. Participants using morally connoted language displayed *decreasing* marginal returns from moral behavior, whereas those with less morally connoted language exhibited *increasing* marginal returns.

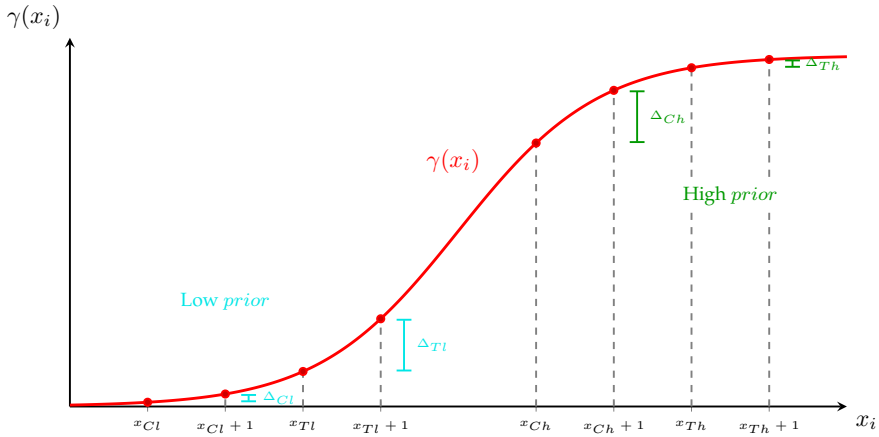
As discussed in Section 3.5.1, language can serve as a proxy for moral sensitivity and prior moral experience. We therefore conjecture that these two groups do not represent fundamentally different types of individuals, but rather individuals at different stages of moral experience. If so, our findings can be integrated into a single, non-monotonic moral satisfaction function $\gamma(\cdot)$, with each group corresponding to a distinct region of its domain.

Participants with low prior moral engagement donate more when morally stimulated, consistent with $\gamma(\cdot)$ being convex at low levels. In the left part of Figure 10, the marginal return from an additional euro donated is lower in the control condition (Δ_{Cl}) than in the treatment (Δ_{Tl}). This suggests that individuals in low-moralization contexts, reflected in their language, experience lower overall moral satisfaction and thus respond positively to moral stimulation, in line with Kraut, 1973.

By contrast, participants with high prior engagement reduce their donations after moral stimulation, consistent with $\gamma(\cdot)$ turning concave. In the right part of the graph, the marginal return is higher under control (Δ_{Ch}) than under treatment (Δ_{Th}), in line with the findings from Benoit Monin and Miller, 2001, where prior moral actions led to a temporary relaxation of subsequent moral behavior.

Together, these patterns suggest an S-shaped $\gamma(\cdot)$: initial moral actions increase satisfaction at an accelerating rate, but at higher engagement, additional actions yield diminishing returns. This framework reconciles evidence that moral labeling can either reinforce consistency (Freedman and Fraser, 1966) or license moral backsliding (Benoit Monin and Miller, 2001).

Figure 10: *Graphical representation of* $\gamma(x_i)$



Notes:
x_{Cl} := moral actions after Phase 1, control, low *prior*.
Δ_{Cl} := marginal moral satisfaction for an additional action, control, low *prior*.
x_{Tl} := moral actions after Phase 1, treatment, low *prior* (control low + 100 donation).
Δ_{Tl} := marginal moral satisfaction for an additional action, treatment, low *prior*.
x_{Ch} := moral actions after Phase 1, control, high *prior*.
Δ_{Ch} := marginal moral satisfaction for an additional action, control, high *prior*.
x_{Th} := moral actions after Phase 1, treatment, high *prior* (control high + 100 donation).
Δ_{Th} := marginal moral satisfaction for an additional action, treatment, high *prior*.

3.6 Literature

We build on a large literature that models *moral satisfaction*, often labeled the "warm glow" of giving, as a hedonic payoff from acting in line with one's moral values. Classic and modern theories embed this motive in different ways, including impure altruism and related formulations in which prosocial acts generate intrinsic utility beyond material outcomes (Andreoni, 1990; Akerlof and Kranton, 2000; Bénabou and Tirole, 2006; Bénabou and Henkel, 2025; Alger and Weibull, 2013; Capraro, Halpern, and Perc, 2024; Minardi, Wang, and Gilboa, 2024). These frameworks rationalize a broad set of behavioral patterns, including moral balancing and self-licensing, by allowing past moral actions to affect the agent's subsequent incentives. However, this literature typically treats warm glow as a reduced-form term and does not connect *the shape* of the moral-satisfaction function to the dynamic patterns of moral behavior observed across repeated decisions.

A complementary empirical literature documents the existence and salience of moral satisfaction in prosocial contexts. Experimental and field evidence shows that individuals often value giving for its own sake, consistent with an additional moral or hedonic component in utility (Andreoni, 1993; Crumpler and P. J. Grossman, 2008; Braaten, 2014; Konow, 2010; Andreoni and Payne, 2011; Tonin and Vlassopoulos, 2014; Abeler, Nosenzo, and Raymond, 2019). Relatedly, several studies design settings in which the researcher can isolate the moral value of actions such as donating from their material consequences, by holding fixed (or removing) the real effects of the transfer (Burum, Nowak, and Hoffman, 2020; Crumpler and P. J. Grossman, 2008). Work in neuroscience further supports this interpretation by linking prosocial choices to reward-related neural responses (Moll et al., 2006; W. T. Harbaugh, Mayr, and Burghart, 2007). Despite this evidence, researchers rarely measure how the moral-satisfaction payoff changes at the margin as individuals repeat morally connoted actions.

Most existing models and empirical implementations therefore impose structure on the functional form. A common assumption is that

moral satisfaction increases with prosocial behavior but exhibits diminishing marginal returns (DellaVigna, List, and Malmendier, 2012; DellaVigna, List, Malmendier, and Rao, 2013). This restriction delivers sharp predictions and tractability, but it also precludes the possibility that marginal moral satisfaction increases for some individuals or in some contexts—a feature that could naturally generate moral consistency rather than moral licensing within a unified preference-based framework.

Recent work begins to estimate the shape of warm glow directly. In particular, Carpenter (2021) uses a field experiment to structurally recover warm-glow curvature and finds meaningful heterogeneity: participants with higher warm glow display increasing preferences with diminishing marginal returns, while others appear motivated primarily by outcome-based incentives. We contribute to this emerging line by focusing explicitly on how heterogeneity in the marginal moral-satisfaction profile maps into different dynamic patterns of behavior (consistency versus licensing) when individuals face repeated, morally framed decisions.

We also contribute to the literature on dynamic moral behavior by providing a unified preference-based mechanism that can generate both moral consistency and moral licensing. Early work documented self-reinforcing patterns of prosociality, through the "foot-in-the-door" effect (Freedman and Fraser, 1966), evidence on repeated helping (Kraut, 1973; Beaman et al., 1983; Burger, 1999), and theories of self-perception (Festinger, 1957; Gawronski and Strack, 2012). More recent research, in contrast, emphasizes moral licensing, whereby an initial good deed reduces subsequent prosociality (Benoit Monin and Miller, 2001; Merritt, Effron, and Benoît Monin, 2010; Jordan, Mullen, and Murnighan, 2011; Brañas-Garza et al., 2013; Ploner and Regner, 2013; Sachdeva, Iliev, and Medin, 2009). Reviews stress that both patterns coexist in the data, yet few studies examine them jointly and theoretical frameworks reconciling them remain limited (Mullen and Benoît Monin, 2016; Ferguson et al., 2024). Our approach speaks directly to this gap by modeling a single moral-satisfaction channel capable of delivering either dynamic pattern across individuals.

3.7 Conclusion

This paper develops and tests a unified framework for understanding dynamic moral behavior. We propose that heterogeneity in the curvature of moral satisfaction, the hedonic utility derived from acting in line with one's values, explains why some individuals behave consistently across moral decisions while others exhibit moral licensing. Using an online experiment that isolates moral satisfaction from material outcomes, and a language-based measure of moral engagement, we identify systematic variation in the shape of this moral-satisfaction function. Participants displaying higher moral engagement show decreasing marginal satisfaction, consistent with moral licensing, whereas less engaged individuals exhibit increasing marginal satisfaction, consistent with moral consistency.

These results provide microfoundations for the coexistence of moral consistency and licensing within a single preference-based model. They suggest that moral behavior is shaped not only by external payoffs but also by the internal dynamics of hedonic reinforcement. When moral satisfaction is concave, additional good deeds yield diminishing returns, encouraging moral substitution; conversely, when moral satisfaction is convex, moral actions reinforce one another, sustaining prosocial consistency. The framework also clarifies why previous studies have produced mixed results, as they often confounded outcome-based with norm-based manipulations.

An additional dimension of our analysis concerns the coexistence of multiple ethical domains. Our findings have important policy implications: moral actions in one domain (such as environmentally responsible behavior) may crowd out prosocial engagement in another, such as blood donation, due to the concavity of moral satisfaction. This substitution effect underscores the need to account for cross-domain moral dynamics when designing interventions aimed at promoting sustained prosocial behavior. Hence, interventions that evoke moral satisfaction in one domain may lead to reduced engagement in others. Firms can exploit this substitutability by "selling" moral satisfaction through sym-

bolic gestures (such as greenwashing) that generate private moral utility without genuine welfare gains.

From a welfare perspective, recognizing moral satisfaction as an endogenous, limited moral resource reframes how policymakers and organizations should design moral incentives. Effective interventions must consider individuals' position along the moral-satisfaction curve: nudges that strengthen prosocial engagement for some may inadvertently induce licensing for others. A socially optimal allocation of moral incentives should therefore target domains with the highest welfare returns and guard against the dilution of moral motivation through purely hedonic or symbolic appeals.

Finally, this work has focused primarily on the positive aspects of moral experience, specifically the satisfaction derived from doing the right thing. A natural next step is to examine the loss domain: the negative moral utility associated with immoral or norm-deviant actions. We expect this domain to exhibit distinctive dynamics, potentially asymmetric with those identified in the gain domain. Exploring this side of moral utility would also illuminate the mechanisms underlying moral balancing, cleansing, and the accumulation of moral credits (Nisan, 1991; Merritt, Effron, and Benoît Monin, 2010; Brañas-Garza et al., 2013; Fanghella and Thøgersen, 2022). Future research should further clarify the conceptual boundary between good, bad, and neutral actions, which we hypothesize depends on the degree of activity or passivity each entails. Prior work suggests that moral and prosocial behaviors differ substantially depending on whether they require active engagement or passive restraint (Teper and Inzlicht, 2011; Adelman, Verkuyten, and Yogeeswaran, 2022). This distinction determines whether individuals perceive the same behavior as fulfilling a prescriptive norm or as avoiding a violation of a proscriptive one. For instance, refraining from lying can be interpreted either as doing good by upholding honesty or as not doing bad by abstaining from deceit, and the dominant framing may crucially influence moral evaluation and motivation. Integrating these insights with the present framework could help distinguish how moral satisfaction operates in active versus passive moral contexts and ultimately enable a

unified account of moral behavior that encompasses both moral gains and moral losses within a single dynamic preference model.

3.8 Appendix: Theoretical Specifications

This appendix provides a detailed account of the theoretical underpinnings of our study. We first clarify the notion of ethical domains and how they frame morally relevant decisions in everyday contexts. Next, we present the game-theoretic considerations that inform our modeling approach, highlighting how moral valuations interact with both individual choices and strategic or uncertain environments. The aim is to make explicit the assumptions and formalizations that underpin our analysis, linking moral theory, normative beliefs, and decision-making in a coherent framework.

3.8.1 Ethical Domains

In this paper, we refer to ethical domains (or more generally, domains)in the context of morally charged decision-making. It is important to clarify that by domains, we mean concrete, everyday situations where individuals face moral choices. Examples include environmentally responsible behaviors, helping less fortunate individuals, supporting social justice initiatives, assisting victims of natural disasters, or providing healthcare to elderly people. Each of these situations represents a specific context in which people make morally relevant decisions, guided by their personal normative beliefs about what actions are morally appropriate.

The term ethical domain may evoke the five foundations proposed by Moral Foundations Theory (J. Graham et al., 2013). However, it is essential to distinguish between the concept of moral domains, as used in this paper, and that of moral foundations. Actions can fall under the same foundation yet belong to different domains, and vice versa.

To illustrate with a simple example: donating to a charity that rescues migrants in the Mediterranean may fall under the *Care* foundation and be associated with the domain of Migration. Alternatively, donating to a charity that supports communities in migrants' countries of origin, emphasizing the goal of reducing migration by helping people "at home", would belong to the same domain (Migration) but could be associated with other foundations, such as *Authority* or *Purity*.

3.8.2 Game-Theoretic Considerations

In this Section, we want to clarify the interpretation of our modeling choices and how they extend to settings with strategic complexities.

Firstly, in the very simple decision theory setting, there is a full coincidence between the action selected and the outcome realized. It is easy to see that this is not the case in a strategic interaction setting, where the actions of other players influence the outcome. But this is also true when the outcome is uncertain and not fully dependent on the decision maker's choice. It is crucial to underline that the moral premium does not depend on the outcome, but rather on the action selected. This means that the argument of the function $u()$ is the realized outcome, in terms of payoffs. Conversely, the argument of the function γ is a combination of the chosen action, its framing, and its evaluation according to the decision maker's personal normative belief[16].

For a full descriptive formalization, we can explicitly define $\pi : X \times Y \rightarrow \mathbb{R}^n$ as a function that tells the actual bundle the individual consumes given the chosen bundle $x \in X$ and the choice of others $y \in Y$.This represents a generalization of this model from this simple deterministic decision theory setting to a possibly random or strategic framework.

Given the strategic specification s, the labeling specification l and the personal norm described by $D^{n\times K}$, the consumer maximizes

$$U(x) = u(\pi(x,s)) + \gamma(\cdot)\mathbf{1}_{\{\sum_{k=1}^{K}\sum_{j=1}^{n} x_j \cdot l_{jk} \cdot d_{jk} > 0\}}$$

where $\sum_{k=1}^{K}\sum_{j=1}^{n} x_j \cdot l_{jk} \cdot d_{jk} > 0$ if and only if there exists a good j, presented with a morally salient label $l_{jk} = 1$ and fulfilling the principle (d_{kj}), that is consumed at a positive quantity in x under label l.

In our simplified decision theory setting, from now on, $\pi(x,s) = x$, but it is crucial to understand that the outcome-based component of the utility function $u(\cdots)$ depends on π, while the norm-based one γ does not.

[16]Referring to the formalization by Alger and Weibull, 2013 and Leeuwen and Alger, 2024, $u_i(x)$ represents their $u(x,y)$, while $\gamma(\cdot)$ is a slightly more general representation of the moral concern that they call $u(x,x)$.

3.9 Appendix: Design

This appendix provides a detailed overview of the experimental interface and procedures used in the study. Participants interacted with a sequence of screens designed to guide them through the experiment, ensure comprehension of key instructions, elicit their choices and beliefs, and record their motivations. Each figure illustrates a specific stage of the experiment, from the introduction of the payment procedure to the donation decision, the elicitation of normative beliefs, and the final feedback. This detailed documentation aims to enhance the transparency and reproducibility of the study.

3.9.1 Description of the experimental structure and payment procedures

After reading and approving the consent form, participants proceed to the initial instructions. On this screen, we provide detailed information about the structure of the experiment and the payment procedure (Figure 11). Participants are informed that, at the end of the study, we randomly select 1 out of 100 participants. For the selected participant, her decisions determine both her personal monetary payoff and the amount donated to the Italian Red Cross.

All remaining participants receive a fixed participation fee of €0.50. The decisions of non-selected participants do not affect either the payoff of the selected participant or the donation to the Italian Red Cross.

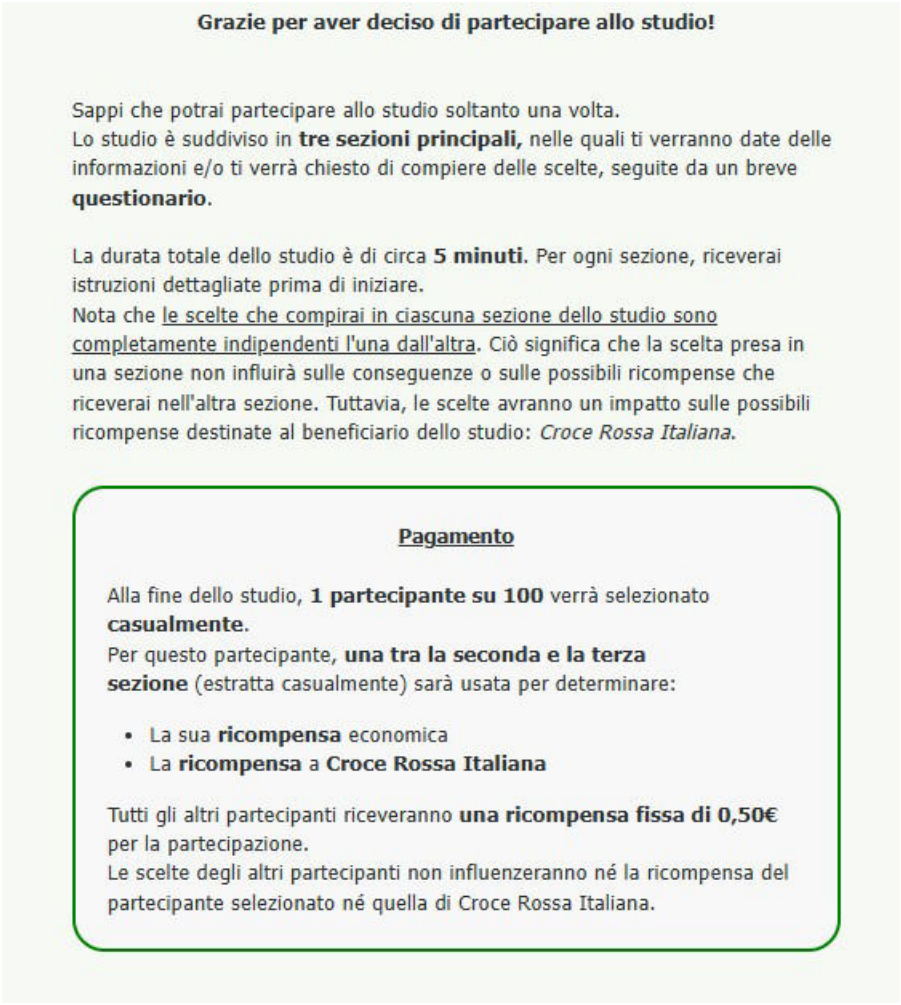


Figure 11: *Description of the experimental structure and the payment procedure*

3.9.2 Description of the activities of the Italian Red Cross

After presenting the payment procedure and before starting the main experimental tasks, participants view a screen introducing the Italian Red Cross (IRC), its areas of operation, and the implementation of donations

(Figure 12). Specifically, participants are informed that, during the study, they will have the opportunity to make a donation to the Italian Red Cross.

The screen explains that the Red Cross is a volunteer-based organization dedicated to providing healthcare and social assistance in both peacetime and times of conflict. It operates through national branches, such as the Italian Red Cross or the Red Cross more broadly, which allows the experiment to be replicated across countries without compromising the credibility or perceived proximity of the charity.

As a well-established and widely respected institution, the Red Cross operates in multiple sectors. In the context of this study, we interpret each sector as a distinct *moral domain*. At this stage, participants are also informed that, in line with standard Red Cross procedures, donors cannot choose the specific destination of their donation. Here the link to donate to Italian Red Cross.

Figure 12: *Description of the charity*

3.9.3 Comprehension Check

Next, participants completed a comprehension check to ensure they understood that they had no control over how the donation would be allocated (Figure 13)[17]

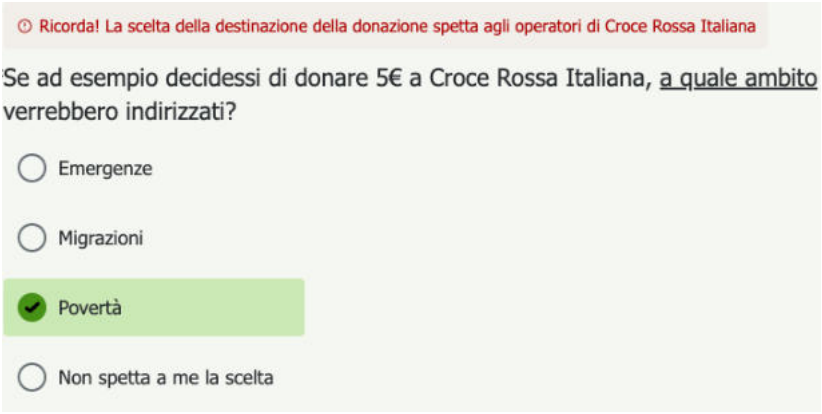


Figure 13: *Comprehension check*

3.9.4 First Phase

Participants then enter the first main part of the experiment. Participants assigned to the Control condition are simply informed that a 100 euro donation has been made to the Italian Red Cross, and reminded that the allocation of the funds across activities will be determined by the charity's operators.

Participants assigned to the Same Domain and Cross Domain conditions are instead informed that, thanks to their participation in the study, a 100 euro donation has been made to the Italian Red Cross. They are reminded that they cannot choose the specific destination of the donation. They are then asked to select a statement that communicates the

[17] Participants could proceed only if they selected *It is not my choice"*. Selecting an alternative triggers an error message: *This question is meant to verify that you understand how donations are used: remember, it is the Italian Red Cross operators who decide how to allocate donations."*

meaning they wish to attribute to this donation.

After selecting a statement, participants receive a personalized pop-up acknowledging their choice:

> *"Thank you! Now we know that [selected domain] is important to you."*

The pop-up also displays the image corresponding to the selected moral domain (Figure 14).

Figure 14: *Popup message example for the domain "Health"*

3.9.5 Second Phase

Participants then enter the second main part of the experiment. They are informed that they are endowed with 100 euro and can use this amount to make a donation to the Italian Red Cross. Before making the donation decision, participants are asked to select or, depending on the treatment, to read again a statement that authorizes the donation and communicates the meaning they wish to attach to it.

The statement-selection task varies by experimental condition. Participants assigned to the Same Domain condition read again the statement they selected in the first part of the experiment. Participants in the Cross Domain condition select one statement from the three remaining options that were not chosen in the first part. Participants in the Control condition select a statement for the first time from the full set of four available statements.

Figure 15 displays the statement-selection interface used in the second phase of the experiment for participants in the Control condition.

Figure 15: *Second Phase: Statement selection interface*

After selecting a statement, participants used a slider to allocate €100 between themselves and the Italian Red Cross (Figure 16).

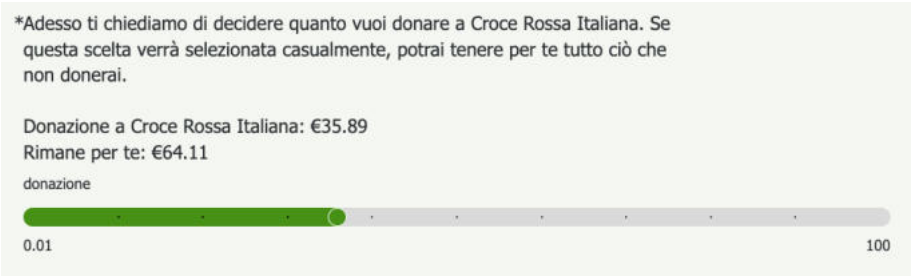


Figure 16: *Second Phase: Donation interface*

3.9.6 Personal and Social Normative Belief Elicitation

Following the main task, we elicited participants' personal normative beliefs, i.e., what they think they *should* do. This allowed us to control for underlying inclinations toward the five moral domains explored earlier. Following Bicchieri (2016), participants indicated their agreement with statements such as:

> "Achieving fair access to health services for all is a goal that I feel it is my duty to help pursue."

Responses were recorded on a 4-point Likert scale (*Strongly Disagree* to *Strongly Agree*) to capture subtle differences in normative convictions. To reduce order effects, the questions were presented in randomized order (Figure 17).

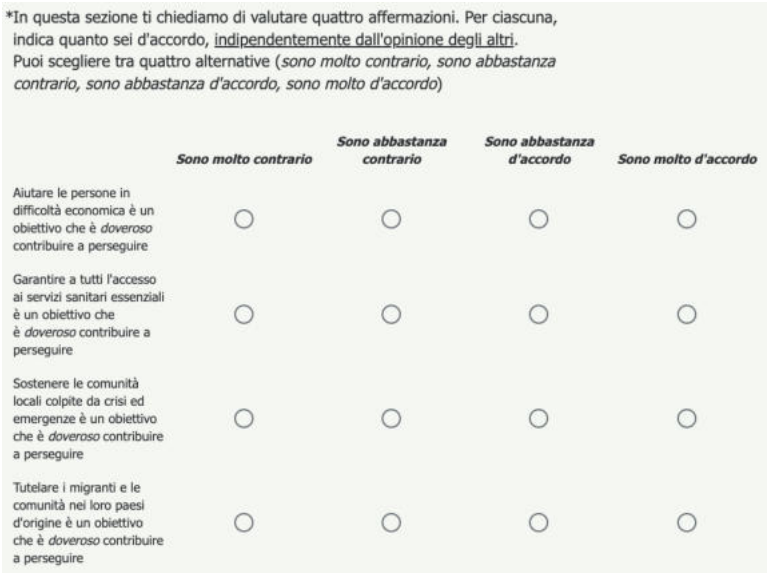
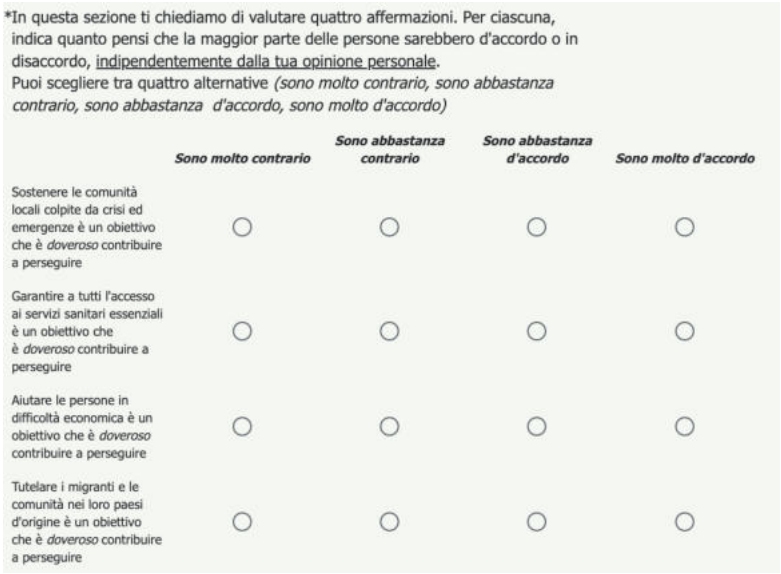


Figure 17: *Personal normative belief elicitation*

From a theoretical perspective, these questions ensure that d_{jk} takes the value 1 for the action of donating to the Italian Red Cross, within the specified domains.

We also measured participants' social normative expectations, i.e., their perception of what others think one *should* do (Bašić and Verrina, 2024; Bicchieri, 2016) (Figure 18). Controlling for these perceptions is important given the role of social norms in shaping personal values.


*In questa sezione ti chiediamo di valutare quattro affermazioni. Per ciascuna, indica quanto pensi che la maggior parte delle persone sarebbero d'accordo o in disaccordo, indipendentemente dalla tua opinione personale.
Puoi scegliere tra quattro alternative *(sono molto contrario, sono abbastanza contrario, sono abbastanza d'accordo, sono molto d'accordo)*

	Sono molto contrario	***Sono abbastanza contrario***	***Sono abbastanza d'accordo***	***Sono molto d'accordo***
Sostenere le comunità locali colpite da crisi ed emergenze è un obiettivo che è *doveroso* contribuire a perseguire	○	○	○	○
Garantire a tutti l'accesso ai servizi sanitari essenziali è un obiettivo che è *doveroso* contribuire a perseguire	○	○	○	○
Aiutare le persone in difficoltà economica è un obiettivo che è *doveroso* contribuire a perseguire	○	○	○	○
Tutelare i migranti e le comunità nei loro paesi d'origine è un obiettivo che è *doveroso* contribuire a perseguire	○	○	○	○

Figure 18: *Social normative belief elicitation*

3.9.7 Support for the Italian Red Cross

Participants' support for volunteer organizations in general, and the Italian Red Cross in particular, was measured on a 1–10 scale:

- 1 = Do not support the work of voluntary organizations (Italian Red Cross) at all"

- 10 = Totally support the work of voluntary organizations (Italian Red Cross)"

This measure serves both as a control for prior attitudes and as a proxy for prosociality (Figure 19).

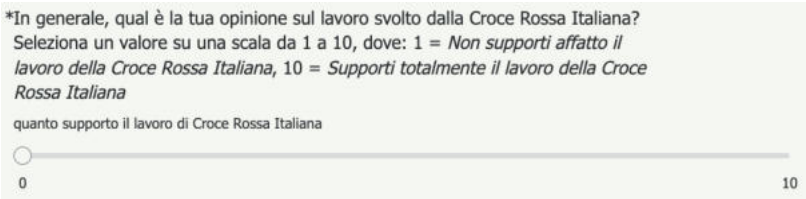


Figure 19: *Support for Italian Red Cross*

Finally, participants' motivations for their choices during the experiment were collected via an open-ended question (Figure 20).

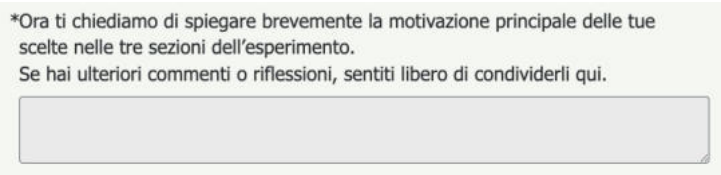


Figure 20: *Motivation elicitation*

3.9.8 Dealing with Image-induced Bias

Given the extensive literature suggesting that exposure to images can bias responses, by triggering emotional reactions or affecting perceptions of salience and urgency Small, Loewenstein, and Slovic, 2013; Shr et al., 2019, we addressed this potential concern by designing two versions of the survey, identical in textual content but differing in the images shown. Specifically, each domain is associated with two different images, one per survey version. In selecting the images, we ensured that both depicted Red Cross operators, in recognizable uniforms, actively engaged in the relevant activity for that domain. This consistency helps control for visual salience while still allowing us to test for possible framing effects.

Participants are randomly assigned to one of the two survey versions. By comparing responses across versions, we can assess whether image exposure significantly alters participants' choices or perceptions. This provides a robustness check against potential confounding effects due to

visual framing, helping to ensure that any observed treatment effects are not artifacts of the imagery used.

In our analysis, we will include a set of robustness checks where we estimate the main models separately by survey version and test for systematic differences in responses. If no significant differences emerge, we interpret this as evidence that our findings are not driven by image-induced bias.

3.10 Appendix: Analysis

This appendix provides a detailed examination of the experimental results, focusing on the heterogeneity of donation behavior and the moral content of participants' motivations. We first analyze the distribution of donations across treatment groups to highlight patterns consistent with distinct normative anchors and the influence of personal norms. Next, we investigate participants' open-ended responses, using both AI-based classification (LLAMA) and a keyword-based approach, to determine whether motivations reflect explicit moral reasoning. Together, these analyses aim to deepen our understanding of how normative beliefs shape behavior and to provide robustness checks for the main experimental findings.

3.10.1 Heterogeneity of Norms: equal split and full generosity

Plotting the Cumulative Distribution Function (CDF) and the Probability Density Function (PDF) provides further insight: donations in the control group tend to be lower and more dispersed than in the cross-domain and same-domain groups, which instead have higher picks in the common norms 50 (equal split) and 100 (pure generosity). This is also confirmed by the higher variance (Variance of Control: 993.1988; Variance of Cross Domain: 861.3138; Variance of Same Domain: 881.1706).

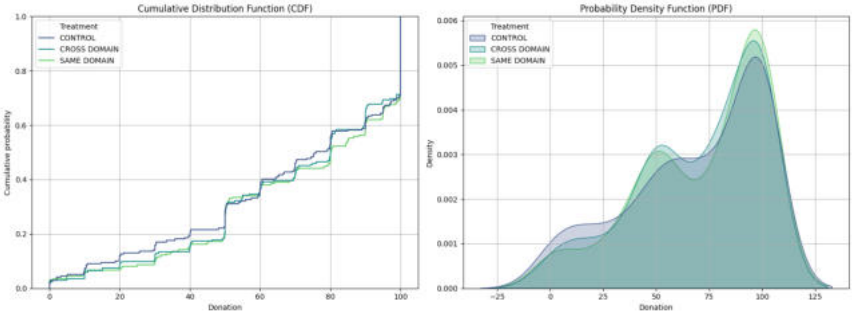


Figure 21: *CDF and PDF by treatment*

Given the possibility of heterogeneity in personal norms among participants (with some anchoring at 50 and others at 100), a factor not explicitly addressed in our initial analysis, we should consider the following mechanism. For participants in the control group who would have donated less than 50, the treatments act as a boost, increasing their donations. In contrast, for those in the control group who would have donated more than 50, the treatment boost is ineffective, as their personal norm was already satisfied. This mechanism explains why the probability of observing a donation below 50 is higher in the control group than in the treatment groups. When considering donations above or equal to 50, this distinction is no longer significant.

Both the distribution of donations and participants' qualitative comments suggest that individuals tend to anchor their donation behavior around two distinct normative reference points: approximately 50 or 100. For example:

- Subject 18, who donated 49.82 in the cross-domain treatment, explained:

  > *La motivazione principale è allineare le donazioni a quelle che secondo me sono le priorità e dividere equamente la donazione tra me e Croce Rossa.*
  >
  > *(Transl. The main motivation is to align the donations with what I consider priorities and to split the donation equally between myself and the Red Cross.)*

When we isolate donations below 50, a range in which the personal norm suggests that individuals "should" donate more, we observe that both the distribution and the average donation differ meaningfully between the control and treatment groups. This pattern is consistent with a convex shape of the function $\gamma(\cdot)$ governing donation behavior.

The lack of a significant difference when considering the entire sample can be explained by the presence of two opposing normative anchors. For donations above 50, the norm at 50 is already exceeded and does not increase contributions, while the norm at 100 motivates some partic-

ipants to give more. These different effects mitigate each other, resulting in an overall insignificant difference in aggregate donations.

Table 11: Mann–Whitney U test results for different value ranges

Sample	Comparison	U Statistic	P-value
Full sample	Control vs. Cross Domain	19782.0	0.783
	Control vs. Same Domain	18866.5	0.513
	Cross Domain vs. Same Domain	19399.5	0.662
≤ 50	Control vs. Cross Domain	862.0	**0.128**
	Control vs. Same Domain	825.0	**0.098**
	Cross Domain vs. Same Domain	963.0	0.828
> 50	Control vs. Cross Domain	12080.0	0.847
	Control vs. Same Domain	11353.5	0.714
	Cross Domain vs. Same Domain	11528.5	0.530

3.10.2 Text Analysis

LLAMA Classification To classify participants' motivations as morally connoted or not, we employed LLAMA (Touvron et al., 2023), an AI language classifier. Manual spot-checks confirmed that classifications aligned well with human judgment. The model was prompted with the following instruction:

```
Sei un classificatore linguistico esperto.
Il tuo compito è stabilire in modo rigoroso se una frase è
''moralmente connotata'' oppure ''non moralmente connotata''.

Definizione di ''moralmente connotata'':
La frase rivela esplicitamente che la persona
ha riflettuto sui propri valori morali,
principi etici, senso del dovere o convinzioni
personali per giustificare la risposta o la scelta.
Può essere positiva, neutra o negativa: anche una
critica, un disaccordo o un'opinione contraria
possono essere moralmente connotati solo se
richiamano esplicitamente valori o principi morali.
Sono presenti riferimenti espliciti e diretti
a concetti come: valori, principi morali, etica,
coscienza, giustizia, dovere, bene/male (in senso morale),
ciò che è giusto/sbagliato (in senso etico),
```

```
o ad una valutazione etica soggettiva e basata sui
propri standard morali. La decisione o opinione è
motivata da ciò che la persona considera giusto o
sbagliato secondo i propri standard morali,
non solo per convenienza, efficacia o preferenza pratica.

Definizione di ``non moralmente connotata'':
La frase non mostra esplicitamente un
ragionamento legato ai valori morali personali
o a un senso del dovere morale.
Può contenere opinioni, preferenze, giudizi
di valore (non morali), considerazioni pratiche,
strategiche, descrittive, logiche o basate
sull'efficienza, ma senza menzionare esplicitamente
valori o principi morali. Frasi che descrivono
fatti, esprimono preferenze personali non legate
alla moralità, o valutano l'efficacia/convenienza
sono generalmente non moralmente connotate.
```

The model output is binary: 1 if a sentence is morally connoted, 0 otherwise.

Keyword Classification As a robustness check, we implemented a simpler keyword-based index. A sentence is classified as *morally connoted* if it contains at least one of the following terms:

```
moral_keywords = [
    'solidarietà', 'giustizia', 'empatia', 'compassione',
    'equità', 'responsabilità', 'aiuto', 'supporto',
    'bene', 'diritti', 'dignità','rispetto', 'umanità',
    'persone in difficoltà', 'bisogno', 'aiutare',
    'vulnerabili', 'sofferenza', 'salvare vite',
    'uguaglianza', 'eguale', 'uguale', 'giusto',
    'mia parte', 'propria parte', 'impegno',
    'contribuire', 'contributo', 'morale', 'moralmente'
]
```

Robustness Results The two approaches produce qualitatively consistent results, though the keyword index yields weaker significance, as expected given its lower precision. Also for the keyword index, chi-squared tests show no significant difference in distribution between control and treatment ($\chi^2 = 16.91$, $p = 0.153$). Within the morally connoted group, treatment effects are not significant in either a t-test ($t = 1.23$,

$p = 0.220$) or a Mann–Whitney test ($U = 5305$, $p = 0.116$). For the non-moral group, the Mann–Whitney test shows borderline significance ($U = 4432$, $p = 0.053$), while the two-sample t-test indicates a significant difference ($t = -2.06$, $p = 0.041$).

Overall, results align with the LLAMA classification, though statistical precision is lower under the keyword approach.

Examples of divergence between methods To better understand the differences between the two indices, we examined cases where the LLAMA and keyword classifications diverge. The discrepancies are generally systematic rather than random. The keyword method tends to over-classify sentences as moral whenever terms such as *solidarietà* or *aiuto* appear, even if the reasoning is not explicitly moral. For example:

- "Ritengo importante la solidarietà per il benessere della società."
  "I consider solidarity important for the well-being of society."
- "Lavoro nella sanità. Il contributo della Croce Rossa è fondamentale."
  "I work in healthcare. The contribution of the Red Cross is fundamental."
- "Le mie scelte sono basate su quelli che ritengo essere gli ambiti in cui c'è più bisogno di aiuto economico e umanitario."
  "My choices are based on what I believe are the areas with the greatest need for economic and humanitarian help."

Here, the keyword index classifies as moral (due to words like *solidarietà* and *aiuto*), but LLAMA does not, as the sentences describe practical or institutional considerations rather than explicit moral reasoning.

Conversely, LLAMA identifies as moral some sentences that contain no keywords but make clear reference to moral values or principles:

- "Penso che tutelare le persone più deboli e fragili e con meno possibilità sia la cosa più importante a prescindere."
  "I think that protecting the weakest and most vulnerable people, those with fewer opportunities, is the most important thing regardless."

- "Ho scelto l'accesso ai servizi sanitari essenziali perché ritengo che tutti, davvero tutti, debbano avere la possibilità di essere curati."
  "I chose access to essential health services because I believe that everyone, truly everyone, should have the possibility to be treated."

- "Ho scelto in base a quello che ritengo più importante nella mia scala di valori."
  "I chose based on what I consider most important in my scale of values."

These examples illustrate why LLAMA provides a more precise measure: it captures the underlying moral reasoning even in the absence of explicit keywords, while avoiding false positives when surface terms are used in a purely descriptive sense.

Conclusion

This dissertation has explored how behavioral factors shape economic interactions across social, informational, and moral domains. By bringing social preferences, behavioral beliefs, and moral motivations into formal economic analysis, it has uncovered mechanisms that remain hidden in more traditional models. Taken together, the three studies show that when individuals are socially embedded, face informational frictions, or act on moral motives, their behavior departs from standard predictions in systematic and interpretable ways.

The theoretical insights developed here also invite empirical testing. A natural next step is to examine the strategic predictions of the models using experimental or observational data. Doing so would enable the evaluation of the robustness of the mechanisms proposed in this dissertation and the assessment of their relevance in real-world contexts, such as redistribution, communication, and moral decision-making.

Overall, this work contributes to a broader understanding of human behavior by showing how concerns about status, communication incentives, and moral motivation jointly influence economic choices. Extending these ideas, both empirically and theoretically, can help build a richer and more realistic foundation for behavioral economics, one that better reflects the complexity of human motives and their implications for social welfare and policy.

Bibliography

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond (2019). "Preferences for Truth-telling". In: *Econometrica* 87.4, pp. 1115–1153. DOI: https://doi.org/10.3982/ECTA14673. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA14673. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14673.

Adelman, Louise, Maykel Verkuyten, and Kumar Yogeeswaran (Aug. 2022). "Distinguishing Active and Passive Outgroup Tolerance: Understanding Its Prevalence and the Role of Moral Concern". In: *Political Psychology* 43.4. Epub 2021 Nov 12, pp. 731–750. DOI: 10.1111/pops.12790.

Adriani, Fabrizio and Silvia Sonderegger (2019). "A theory of esteem based peer pressure". In: *Games and Economic Behavior* 115, pp. 314–335. ISSN: 0899-8256. DOI: https://doi.org/10.1016/j.geb.2019.03.010. URL: https://www.sciencedirect.com/science/article/pii/S0899825619300454.

Akerlof, George A. and Rachel E. Kranton (2000). "Economics and Identity". In: *The Quarterly Journal of Economics* 115.3, pp. 715–753. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/2586894 (visited on 10/21/2025).

Alger, Ingela and Jörgen W. Weibull (2013). "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching". In: *Econometrica* 81.6, pp. 2269–2302. DOI: https://doi.org/10.3982/ECTA10637. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA10637. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10637.

Andreoni, James (1990). "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving". In: *The Economic Journal* 100.401,

pp. 464–477. ISSN: 00130133, 14680297. URL: http://www.jstor.org/stable/2234133 (visited on 03/21/2025).

— (1993). "An Experimental Test of the Public-Goods Crowding-Out Hypothesis". In: *The American Economic Review* 83.5, pp. 1317–1327. ISSN: 00028282. URL: http://www.jstor.org/stable/2117563 (visited on 10/21/2025).

Andreoni, James and A. Abigail Payne (2011). "Is crowding out due entirely to fundraising? Evidence from a panel of charities". In: *Journal of Public Economics* 95.5. Charitable Giving and Fundraising Special Issue, pp. 334–343. ISSN: 0047-2727. DOI: https://doi.org/10.1016/j.jpubeco.2010.11.011. URL: https://www.sciencedirect.com/science/article/pii/S0047272710001714.

Antinyan, Armenak, Gergely Horváth, and Mofei Jia (2019). "Social status competition and the impact of income inequality in evolving social networks: An agent-based model". In: *Journal of Behavioral and Experimental Economics* 79, pp. 53–69. ISSN: 2214-8043. DOI: https://doi.org/10.1016/j.socec.2018.12.008. URL: https://www.sciencedirect.com/science/article/pii/S2214804318301216.

Bagwell, Kyle and Garey Ramey (1991). "Oligopoly limit pricing". In: *The Rand Journal of Economics*, pp. 155–172.

Bagwell, Laurie Simon and B. Douglas Bernheim (1996). "Veblen Effects in a Theory of Conspicuous Consumption". In: *American Economic Review* 86.3, pp. 349–373. ISSN: 00028282. URL: http://www.jstor.org/stable/2118201 (visited on 07/10/2023).

Bašić, Zvonimir and Eugenio Verrina (2024). "Personal norms — and not only social norms — shape economic behavior". In: *Journal of Public Economics* 239, p. 105255. ISSN: 0047-2727. DOI: https://doi.org/10.1016/j.jpubeco.2024.105255. URL: https://www.sciencedirect.com/science/article/pii/S0047272724001919.

Battaglini, Marco (2002). "Multiple Referrals and Multidimensional Cheap Talk". In: *Econometrica* 70.4, pp. 1379–1401. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/3082002 (visited on 10/12/2025).

Battigalli, Pierpaolo, Gary Charness, and Martin Dufwenberg (2013). "Deception: The role of guilt". In: *Journal of Economic Behavior & Organization* 93, pp. 227–232. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2013.03.033. URL: https://www.sciencedirect.com/science/article/pii/S0167268113000784.

Baumberg, Ben (2016). "The stigma of claiming benefits: a quantitative study". In: *Journal of Social Policy* 45.2, pp. 181–199. DOI: 10.1017/S0047279415000525.

Beaman, Arthur et al. (June 1983). "Fifteen Years of Foot-in-the Door Research: A Meta-Analysis". In: *Personality and Social Psychology Bulletin* 9, pp. 181–196. DOI: 10.1177/0146167283092002.

Bellezza, Silvia, Francesca Gino, and Anat Keinan (2014). "The red sneakers effect: Inferring status and competence from signals of nonconformity". In: *Journal of Consumer Research* 41.1, pp. 35–54. DOI: 10.1086/674870. URL: https://doi.org/10.1086/674870.

Bénabou, Roland, Armin Falk, and Luca Henkel (Jan. 2024). "Ends versus Means: Kantians, Utilitarians, and Moral Decisions". In: *NBER Working Papers*. URL: https://ideas.repec.org/p/nbr/nberwo/32073.html.

Bénabou, Roland and Luca Henkel (Sept. 2025). "Identity as Self-Image". In: *NBER Working Paper No. 34297*. NBER Working Paper Series 34297. DOI: 10.3386/w34297. URL: https://www.nber.org/papers/w34297.

Bénabou, Roland and Jean Tirole (Dec. 2006). "Incentives and Prosocial Behavior". In: *American Economic Review* 96.5, pp. 1652–1678. DOI: 10.1257/aer.96.5.1652. URL: https://www.aeaweb.org/articles?id=10.1257/aer.96.5.1652.

Bicchieri, Cristina (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bilancini, Ennio and Leonardo Boncinelli (2012). "Redistribution and the notion of social status". In: *Journal of Public Economics* 96.9, pp. 651–657. ISSN: 0047-2727. DOI: https://doi.org/10.1016/j.jpubeco.2012.05.007. URL: https://www.sciencedirect.com/science/article/pii/S0047272712000461.

— (2014). "Instrumental cardinal concerns for social status in two-sided matching with non-transferable utility". In: *European Economic Review* 67, pp. 174–189.

Bonferroni, Carlo Emilio (1936). "Teoria statistica delle classi e calcolo delle probabilità". Italian. In: *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8, pp. 3–62.

Bourdieu, Pierre (1979). *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, MA: Harvard University Press.

Braaten, Ragnhild Haugli (2014). "Testing deontological warm glow motivation for carbon abatements". In: *Resource and Energy Economics* 38. Cited by: 3, pp. 96–109. DOI: 10.1016/j.reseneeco.2014.

06.003. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84904732235%5C&doi=10.1016%2fj.reseneeco.2014.06.003%5C&partnerID=40%5C&md5=a6d61bcaba290ef7a400b48a3f090c65.

Brañas-Garza, Pablo et al. (2013). "Moral Cleansing and Moral Licenses: Experimental Evidence". In: *Economics and Philosophy* 29.2, pp. 199–212. DOI: 10.1017/S0266267113000199.

Burger, Jerry M. (1999). "The Foot-in-the-Door Compliance Procedure: A Multiple-Process Analysis and Review". In: *Personality and Social Psychology Review* 3.4, pp. 303–325. ISSN: 1088-8683. DOI: 10.1207/s15327957pspr0304_2. URL: https://doi.org/10.1207/s15327957pspr0304_2.

Burum, Bethany, Martin Nowak, and Moshe Hoffman (Dec. 2020). "An evolutionary explanation for ineffective altruism". In: *Nature Human Behaviour* 4, pp. 1–13. DOI: 10.1038/s41562-020-00950-4.

Capraro, Valerio, Joseph Y. Halpern, and Matjaž Perc (Mar. 2024). "From Outcome-Based to Language-Based Preferences". In: *Journal of Economic Literature* 62.1, pp. 115–54. DOI: 10.1257/jel.20221613. URL: https://www.aeaweb.org/articles?id=10.1257/jel.20221613.

Carlsson, Hans and Eric van Damme (1993). "Global Games and Equilibrium Selection". In: *Econometrica* 61.5, pp. 989–1018. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2951491 (visited on 10/10/2025).

Carpenter, Jeffrey (2021). "The shape of warm glow: Field experimental evidence from a fundraiser". In: *Journal of Economic Behavior & Organization* 191, pp. 555–574. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2021.09.020. URL: https://www.sciencedirect.com/science/article/pii/S0167268121003991.

Causa, Orsetta and Mikkel Hermansen (July 2019). *Income Redistribution through Taxes and Transfers across OECD Countries*. OECD Economics Department Working Papers 1453. Unclassified English. Cancels and replaces the same document of 15 December 2017. Paris: OECD. URL: https://www.oecd.org/eco/workingpapers/.

Charles, Kerwin Kofi, Erik Hurst, and Nikolai Roussanov (2009). "Conspicuous Consumption and Race". In: *The Quarterly Journal of Economics* 124.2, pp. 425–467. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/40506236 (visited on 01/09/2026).

Chen, Chia-Hui, Junichiro Ishida, and Wing Suen (2022). "Signaling Under Double-Crossing Preferences". In: *Econometrica* 90.3, pp. 1225–

1260. DOI: https://doi.org/10.3982/ECTA19210. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19210. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19210.

Chen, Ying (2011). "Perturbed communication games with honest senders and naive receivers". In: *Journal of Economic Theory* 146.2, pp. 401–424.

Chen, Ying, Navin Kartik, and Joel Sobel (2008). "Selecting Cheap-Talk Equilibria". In: *Econometrica* 76.1, pp. 117–136. DOI: 10.1111/j.0012-9682.2008.00820.x. URL: https://doi.org/10.1111/j.0012-9682.2008.00820.x.

Cho, In-Koo and David M. Kreps (1987). "Signaling Games and Stable Equilibria". In: *Quarterly Journal of Economics* 102.2, pp. 179–221. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/1885060 (visited on 05/06/2022).

Crawford, Vincent P and Joel Sobel (1982). "Strategic information transmission". In: *Econometrica: Journal of the Econometric Society*, pp. 1431–1451.

Crumpler, Heidi and Philip J. Grossman (2008). "An experimental test of warm glow giving". In: *Journal of Public Economics* 92.5, pp. 1011–1021. ISSN: 0047-2727. DOI: https://doi.org/10.1016/j.jpubeco.2007.12.014. URL: https://www.sciencedirect.com/science/article/pii/S0047272708000029.

Currid-Halkett, Elizabeth (2019). "The Sum of Small Things: A Theory of the Aspirational Class". In: *American Journal of Sociology* 124.5, pp. 1566–1568. DOI: 10.1086/701691. eprint: https://doi.org/10.1086/701691. URL: https://doi.org/10.1086/701691.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier (2012). "Testing for Altruism and Social Pressure in Charitable Giving". In: *The Quarterly Journal of Economics* 127.1, pp. 1–56. URL: https://ideas.repec.org/a/oup/qjecon/v127y2012i1p1-56.html.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao (2013). "The Importance of Being Marginal: Gender Differences in Generosity". In: *The American Economic Review* 103.3, pp. 586–590. ISSN: 00028282. URL: http://www.jstor.org/stable/23469798 (visited on 10/21/2025).

Emons, Winand and Claude Fluet (2009). "Accuracy versus falsification costs: The optimal amount of evidence under different procedures". In: *The Journal of Law, Economics, & Organization* 25.1, pp. 134–156.

Fanghella, Valeria and John Thøgersen (2022). "Experimental evidence of moral cleansing in the interpersonal and environmental domains".

In: *Journal of Behavioral and Experimental Economics* 97, p. 101838. ISSN: 2214-8043. DOI: https://doi.org/10.1016/j.socec.2022.101838. URL: https://www.sciencedirect.com/science/article/pii/S2214804322000143.

Fehr, Ernst and Klaus M. Schmidt (1999). "A Theory of Fairness, Competition, and Cooperation". In: *The Quarterly Journal of Economics* 114.3, pp. 817–868. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/2586885 (visited on 10/30/2025).

Feltovich, Nick, Rick Harbaugh, and Ted To (Dec. 2002). "Too Cool for School? Signalling and Countersignalling". In: *RAND Journal of Economics* 33, pp. 630–649. DOI: 10.2307/3087478.

Ferguson, Rose et al. (2024). "Influences of past moral behavior on future behavior: A review of sequential moral behavior studies using meta-analytic techniques". In: *Psychological Bulletin* 150.6, pp. 694–726. ISSN: 0033-2909. DOI: 10.1037/bul0000441.

Ferreira, Francisco HG et al. (2012). *Economic mobility and the rise of the Latin American middle class*. The World Bank. DOI: 10.1596/978-0-8213-9634-6. URL: dx.doi.org.

Festinger, Leon (1957). *A Theory of Cognitive Dissonance*. Paperback ISBN: 0-8047-0911-4. Stanford, CA: Stanford University Press. ISBN: 0-8047-0131-8.

Fong, Christina (2001). "Social preferences, self-interest, and the demand for redistribution". In: *Journal of Public Economics* 82.2, pp. 225–246. URL: https://EconPapers.repec.org/RePEc:eee:pubeco:v:82:y:2001:i:2:p:225-246.

Frank, Robert H. (1985). "The Demand for Unobservable and Other Nonpositional Goods". In: *American Economic Review* 75.1, pp. 101–116. ISSN: 00028282. URL: http://www.jstor.org/stable/1812706 (visited on 04/13/2023).

Freedman, Jonathan L. and Scott C. Fraser (1966). "Compliance without Pressure: The Foot-in-the-Door Technique". In: *Journal of Personality and Social Psychology* 4.2, pp. 195–202. DOI: 10.1037/h0023552.

Friedrichsen, Jana, Tobias König, and Tobias Lausen (June 2020). "Social Status Concerns and the Political Economy of Publicly Provided Private Goods". In: *Economic Journal* 131.633, pp. 220–246. ISSN: 0013-0133. DOI: 10.1093/ej/ueaa076. eprint: https://academic.oup.com/ej/article-pdf/131/633/220/36224961/ueaa076.pdf. URL: https://doi.org/10.1093/ej/ueaa076.

Fudenberg, Drew and Jean Tirole (1991). *Game theory*. MIT press.

Gallice, Andrea and Edoardo Grillo (Nov. 2020). "Economic and Social-Class Voting in a Model of Redistribution with Social Concerns". In: *Journal of the European Economic Association* 18.6, pp. 3140–3172. ISSN: 1542-4766. DOI: 10.1093/jeea/jvz061. eprint: https://academic.oup.com/jeea/article-pdf/18/6/3140/34926570/jvz061.pdf. URL: https://doi.org/10.1093/jeea/jvz061.

Gawronski, Bertram and Fritz Strack, eds. (2012). *Cognitive Consistency: A Fundamental Principle in Social Cognition*. New York: Guilford Press, p. 494. ISBN: 978-1-4625-0526-4.

Gentzkow, Matthew and Jesse M Shapiro (2008). "Competition and Truth in the Market for News". In: *The Journal of Economic Perspectives* 22.2, pp. 133–154.

Ghiglino, Christian and Sanjeev Goyal (2010). "Keeping up with the neighbors: Social interaction in a market economy". In: *Journal of the European Economic Association* 8.1. Cited by: 40; All Open Access, Green Open Access, pp. 90–119. DOI: 10.1162/jeea.2010.8.1.90. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-76349084509%5C&doi=10.1162%2fjeea.2010.8.1.90%5C&partnerID=40%5C&md5=9d447791fe19b00b30be3d1ecba9bd3a.

Gilboa, Itzhak (2009). *Theory of Decision under Uncertainty*. Econometric Society Monographs. Cambridge University Press.

Graham, Cameron and Claudine Grisard (2019). "Rich man, poor man, beggar man, thief: Accounting and the stigma of poverty". In: *Critical Perspectives on Accounting* 59, pp. 32–51. ISSN: 1045-2354. DOI: https://doi.org/10.1016/j.cpa.2018.06.004. URL: https://www.sciencedirect.com/science/article/pii/S1045235418301941

Graham, Jesse et al. (2013). "Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism". In: *Advances in Experimental Social Psychology*. Ed. by Patricia G. Devine and Ashby Plant. Vol. 47. Advances in Experimental Social Psychology. San Diego: Academic Press, pp. 55–130. DOI: 10.1016/B978-0-12-407236-7.00002-4. URL: https://www.sciencedirect.com/science/article/pii/B9780124072367000024.

Grossman, Sanford J and Motty Perry (1986). "Perfect sequential equilibrium". In: *Journal of Economic Theory* 39.1, pp. 97–119. ISSN: 0022-0531. DOI: https://doi.org/10.1016/0022-0531(86)90022-0. URL: https://www.sciencedirect.com/science/article/pii/0022053186900220.

Guadagno, Rosanna E and Robert B Cialdini (2010). "Preference for consistency and social influence: A review of current research findings". In: *Social influence* 5.3, pp. 152–163.

Han, Young Jee, Joseph C. Nunes, and Xavier Drèze (2010). "Signaling Status with Luxury Goods: The Role of Brand Prominence". In: *Journal of Marketing* 74.4, pp. 15–30. DOI: 10.1509/jmkg.74.4.015. eprint: https://doi.org/10.1509/jmkg.74.4.015. URL: https://doi.org/10.1509/jmkg.74.4.015.

Harbaugh, William T., Ulrich Mayr, and Daniel R. Burghart (2007). "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations". In: *Science* 316.5831, pp. 1622–1625. DOI: 10.1126/science.1140738. URL: https://doi.org/10.1126/science.1140738.

Hopkins, Ed (2008). "Inequality, happiness and relative concerns: What actually is their relationship?" In: *Journal of Economic Inequality* 6, pp. 351–372.

— (2023). "Cardinal sins? Conspicuous consumption, cardinal status and inequality". In.

Hopkins, Ed and Tatiana Kornienko (Sept. 2004). "Running to Keep in the Same Place: Consumer Choice as a Game of Status". In: *American Economic Review* 94.4, pp. 1085–1107. DOI: 10.1257/0002828042002705. URL: https://www.aeaweb.org/articles?id=10.1257/0002828042002705.

— (2006). "Inequality and growth in the presence of competition for status". In: *Economics Letters* 93.2, pp. 291–296. ISSN: 0165-1765. DOI: https://doi.org/10.1016/j.econlet.2006.05.017. URL: https://www.sciencedirect.com/science/article/pii/S0165176506001923.

— (Aug. 2010). "Which Inequality? The Inequality of Endowments versus the Inequality of Rewards". In: *American Economic Journal: Microeconomics* 2.3, pp. 106–37. DOI: 10.1257/mic.2.3.106. URL: https://www.aeaweb.org/articles?id=10.1257/mic.2.3.106.

Immorlica, Nicole et al. (Feb. 2017). "Social Status in Networks". In: *American Economic Journal: Microeconomics* 9.1, pp. 1–30. DOI: 10.1257/mic.20160082. URL: https://www.aeaweb.org/articles?id=10.1257/mic.20160082.

Ireland, Norman J (1998). "Status-seeking, income taxation and efficiency". In: *Journal of Public Economics* 70.1, pp. 99–113. ISSN: 0047-2727. DOI: https://doi.org/10.1016/S0047-2727(98)00062-0. URL:

https://www.sciencedirect.com/science/article/pii/S0047272798000620.

Ireland, Norman J. (1994). "On limiting the market for status signals". In: *Journal of Public Economics* 53.1, pp. 91–110. ISSN: 0047-2727. DOI: https://doi.org/10.1016/0047-2727(94)90015-9. URL: https://www.sciencedirect.com/science/article/pii/0047272794900159.

Jordan, Jennifer, Elizabeth Mullen, and J. Murnighan (Mar. 2011). "Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior". In: *Personality & social psychology bulletin* 37, pp. 701–13. DOI: 10.1177/0146167211400208.

Kakhbod, Ali and Uliana Loginova (2023). "When does introducing verifiable communication choices improve welfare?" In: *Journal of Economic Behavior & Organization* 210, pp. 139–162. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2023.03.031. URL: https://www.sciencedirect.com/science/article/pii/S016726812300104X.

Kartik, Navin (2009). "Strategic communication with lying costs". In: *The Review of Economic Studies* 76.4, pp. 1359–1395.

Kartik, Navin, Frances Xu Lee, and Wing Suen (June 2021). "Information Validates the Prior: A Theorem on Bayesian Updating and Applications". In: *American Economic Review: Insights* 3.2, pp. 165–82. DOI: 10.1257/aeri.20200284. URL: https://www.aeaweb.org/articles?id=10.1257/aeri.20200284.

Kartik, Navin, Marco Ottaviani, and Francesco Squintani (2007). "Credulity, lies, and costly talk". In: *Journal of Economic Theory* 134.1, pp. 93–116.

Kennedy, Brendan et al. (2021). "Moral concerns are differentially observable in language". In: *Cognition* 212, p. 104696. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2021.104696. URL: https://www.sciencedirect.com/science/article/pii/S0010027721001153.

Konow, James (2010). "Mixed feelings: Theories of and evidence on giving". In: *Journal of Public Economics* 94.3, pp. 279–297. ISSN: 0047-2727. DOI: https://doi.org/10.1016/j.jpubeco.2009.11.008. URL: https://www.sciencedirect.com/science/article/pii/S0047272709001480.

Kraut, Robert E. (1973). "Effects of social labeling on giving to charity". In: *Journal of Experimental Social Psychology* 9.6, pp. 551–562. ISSN: 0022-1031. DOI: https://doi.org/10.1016/0022-1031(73)

90037-1. URL: https://www.sciencedirect.com/science/article/pii/0022103173900371.

Kreps, David M. and Robert Wilson (1982). "Sequential Equilibria". In: *Econometrica* 50.4, pp. 863–894. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912767 (visited on 09/27/2025).

Krishna, Vijay and John Morgan (2001). "A model of expertise". In: *The Quarterly Journal of Economics* 116.2, pp. 747–775.

Krivo, Lauren J. et al. (May 2013). "Social Isolation of Disadvantage and Advantage: The Reproduction of Inequality in Urban Space". In: *Social Forces* 92.1, pp. 141–164. ISSN: 0037-7732. DOI: 10.1093/sf/sot043. eprint: https://academic.oup.com/sf/article-pdf/92/1/141/6884347/sot043.pdf. URL: https://doi.org/10.1093/sf/sot043.

Laplace, Pierre Simon (1812). *Théorie analytique des probabilités*. The foundational text for modern probability theory, in which the Principle of Insufficient Reason is established as the first principle of probability. Paris: Courcier.

Leeuwen, Boris van and Ingela Alger (2024). "Estimating Social Preferences and Kantian Morality in Strategic Interactions". In: *Journal of Political Economy Microeconomics* 2.4, pp. 665–706. DOI: 10.1086/732125. URL: https://doi.org/10.1086/732125.

Levy, Gilat and Ronny Razin (May 2015). "Preferences over Equality in the Presence of Costly Income Sorting". In: *American Economic Journal: Microeconomics* 7.2, pp. 308–337. URL: https://ideas.repec.org/a/aea/aejmic/v7y2015i2p308-37.html.

Manelli, Alejandro M (1997). "The never-a-weak-best-response test in infinite signaling games". In: *Journal of Economic Theory* 74.1, pp. 152–173.

Merritt, Anna C., Daniel A. Effron, and Benoît Monin (2010). "Moral Self-Licensing: When Being Good Frees Us to Be Bad". In: *Social and Personality Psychology Compass* 4.5, pp. 344–357. DOI: https://doi.org/10.1111/j.1751-9004.2010.00263.x. eprint: https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2010.00263.x. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-9004.2010.00263.x.

Mijs, Jonathan J. B. and Elizabeth L. Roe (2021). "Is America coming apart? Socioeconomic segregation in neighborhoods, schools, workplaces, and social networks, 1970–2020". In: *Sociology Compass* 15.6, e12884. DOI: https://doi.org/10.1111/soc4.12884. eprint:

https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/soc4.12884. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12884.

Milgrom, Paul and John Roberts (1986). "Relying on the information of interested parties". In: *The RAND Journal of Economics*, pp. 18–32.

Minardi, Stefania, Fan Wang, and Itzhak Gilboa (2024). "Consumption of Values". In: *Management Science* 0.0, Forthcoming.

Moll, Jorge et al. (2006). "Human Fronto-Mesolimbic Networks Guide Decisions About Charitable Donation". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.42, pp. 15623–15628. DOI: 10.1073/pnas.0604475103. URL: https://doi.org/10.1073/pnas.0604475103.

Monin, Benoit and Dale T Miller (2001). "Moral credentials and the expression of prejudice." In: *Journal of personality and social psychology* 81.1, p. 33.

Mullen, Elizabeth and Benoît Monin (2016). "Consistency Versus Licensing Effects of Past Moral Behavior". In: *Annual Review of Psychology* 67.Volume 67, 2016, pp. 363–385. ISSN: 1545-2085. DOI: https://doi.org/10.1146/annurev-psych-010213-115120. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-psych-010213-115120.

Nagel, Rosemarie (1995). "Unraveling in Guessing Games: An Experimental Study". In: *The American Economic Review* 85.5, pp. 1313–1326. ISSN: 0002-8282. URL: http://www.jstor.org/stable/2950991 (visited on 10/10/2025).

Nisan, Mordechai (1991). "The Moral Balance Model: Theory and Research Extending Our Understanding of Moral Choice and Deviation". In: *Handbook of Moral Behavior and Development*. Ed. by William M. Kurtines and Jacob L. Gewirtz. Vol. 1. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp. 213–249.

O'Cass, Aron and Hmily McEwen (2004). "Exploring consumer status and conspicuous consumption". In: *Journal of Consumer Behaviour* 4.1, pp. 25–39. DOI: https://doi.org/10.1002/cb.155. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cb.155. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cb.155.

OECD (2019). *Under Pressure: The Squeezed Middle Class*. Paris: OECD Publishing. ISBN: 978-92-64-54283-9. DOI: 10.1787/689afed1-en. URL: https://doi.org/10.1787/689afed1-en.

Ottaviani, Marco and Francesco Squintani (2006). "Naive audience and communication bias". In: *International Journal of Game Theory* 35.1, pp. 129–150.

Peetz, Johanna and Andrea L Howard (2022). "People prefer to diversify across different types of prosocial behaviour". In: *British Journal of Social Psychology* 61.3, pp. 924–939.

Pennebaker, James W (2011). *The secret life of pronouns: What our words say about us*. Penguin Group USA.

Pew Research Center (Dec. 2015). *The American Middle Class Is Losing Ground: No longer the majority and falling behind financially*. Washington, D.C.: Pew Research Center.

Piketty, Thomas (1995). "Social Mobility and Redistributive Politics". In: *Quarterly Journal of Economics* 110.3, pp. 551–584. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/2946692 (visited on 09/27/2022).

Ploner, Matteo and Tobias Regner (2013). "Self-image and moral balancing: An experimental analysis". In: *Journal of Economic Behavior & Organization* 93, pp. 374–383. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2013.03.030. URL: https://www.sciencedirect.com/science/article/pii/S0167268113000759.

Ramezani, A. et al. (Sept. 2024). "Evolution of the Moral Lexicon". In: *Open Mind (Camb.)* 8, pp. 1153–1169. DOI: 10.1162/opmi_a_00164.

Reeves, Richard V., Katherine Guyot, and Eleanor Krause (May 2018). *Defining the middle class: Cash, credentials, or culture*. URL: www.brookings.edu.

Riley, John G. (June 2001). "Silver Signals: Twenty-Five Years of Screening and Signaling". In: *Journal of Economic Literature* 39.2, pp. 432–478. DOI: 10.1257/jel.39.2.432. URL: https://www.aeaweb.org/articles?id=10.1257/jel.39.2.432.

Sachdeva, Sonya, Rumen Iliev, and Douglas L. Medin (2009). "Sinning saints and saintly sinners: the paradox of moral self-regulation". In: *Psychological Science* 20.4, pp. 523–528. ISSN: 0956-7976. DOI: 10.1111/j.1467-9280.2009.02326.x. URL: https://doi.org/10.1111/j.1467-9280.2009.02326.x.

Sanesi, Bianca, Ennio Bilancini, and Leonardo Boncinelli (2024). "A Model for Rational Generosity of the Rich: Status Concern and Poverty Blaming When Social Classes are Partially Segregated". Available at SSRN: https://ssrn.com/abstract=4946789 or http://dx.doi.org/10.2139/ssrn.4946789.

Sanesi, Bianca and Ginevra Del Mastio (2025). "The S-Shape of Moral Satisfaction". Available at SSRN: https://ssrn.com/abstract=5503102 or http://dx.doi.org/10.2139/ssrn.5503102.

Selten, Reinhard (1975). "Reexamination of the perfectness concept for equilibrium points in extensive games". In: *International Journal of Game Theory* 4.1, pp. 25–55. ISSN: 1432-1270. DOI: 10.1007/BF01766400. URL: https://doi.org/10.1007/BF01766400.

Shr, Yau-Huo Jimmy et al. (2019). "How do visual representations influence survey responses? Evidence from a choice experiment on landscape attributes of green infrastructure". In: *Ecological Economics* 156, pp. 375–386.

Simmel, Georg and Claire Jacobson (1965). "The Poor". In: *Social Problems* 13.2, pp. 118–140. ISSN: 00377791, 15338533. URL: http://www.jstor.org/stable/798898 (visited on 10/17/2023).

Small, Deborah A, George Loewenstein, and Paul Slovic (2013). "Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims". In: *The feeling of risk*. Routledge, pp. 51–68.

Smith, Adam (None 1759). *The Theory of Moral Sentiments*. None. Vol. None. History of Economic Thought Books smith1759. McMaster University Archive for the History of Economic Thought. DOI: None. URL: https://ideas.repec.org/b/hay/hetboo/smith1759.html.

Spence, Michael (1973a). "Job Market Signaling". In: *Quarterly Journal of Economics* 87.3, pp. 355–374. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/1882010 (visited on 05/06/2022).

— (1973b). "Job Market Signaling". In: *The Quarterly Journal of Economics* 87.3, pp. 355–374. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/1882010.

Stahl, Dale O. and Paul W. Wilson (1995). "On Players' Models of Other Players: Theory and Experimental Evidence". In: *Games and Economic Behavior* 10.1, pp. 218–254. ISSN: 0899-8256. DOI: https://doi.org/10.1006/game.1995.1031. URL: https://www.sciencedirect.com/science/article/pii/S0899825685710317.

Steele, Liza G., Joseph Nathan Cohen, and Joseph R. van der Naald (2022). "Wealth, Income, and Preferences for Redistribution: Evidence from 30 countries". In: *Social Science Research* 108, p. 102746. ISSN: 0049-089X. DOI: https://doi.org/10.1016/j.ssresearch.2022.102746. URL: https://www.sciencedirect.com/science/article/pii/S0049089X22000527.

Teper, Rimma and Michael Inzlicht (2011). "Active Transgressions and Moral Elusions: Action Framing Influences Moral Behavior". In: *Social Psychological and Personality Science* 2.3, pp. 284–288. DOI: 10.1177/1948550610389338. eprint: https://doi.org/10.1177/1948550610389338. URL: https://doi.org/10.1177/1948550610389338.

Toccafondi, Niccolò, Bianca Sanesi, and Sibilla Di Guida (2025). "The Demand Effects of Real-Time ER Congestion Disclosure". Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5591737.

Tonin, Mirco and Michael Vlassopoulos (2014). "An Experimental Investigation of Intrinsic Motivations for Giving". In: *Theory and Decision* 76.1, pp. 47–67. DOI: 10.1007/s11238-013-9360-9. URL: https://doi.org/10.1007/s11238-013-9360-9.

Touvron, Hugo et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971.

Vaccari, Federico (2020). "Influential News and Policy-making". In: *Available at SSRN 3602693*.

— (2023a). "Competition in costly talk". In: *Journal of Economic Theory* 213, p. 105740. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2023.105740. URL: https://www.sciencedirect.com/science/article/pii/S0022053123001369.

— (2023b). "Influential news and policy-making". In: *Economic Theory* 76.4, pp. 1363–1418. DOI: 10.1007/s00199-023-01499-9.

Veblen, Thorstein (1899). *The Theory of the Leisure Class*. McMaster University Archive for the History of Economic Thought. URL: https://EconPapers.repec.org/RePEc:hay:hetboo:veblen1899.

Vida, Peter, Takakazu Honryo, and Helmuts Azacis (2022). *Strong Forward Induction in Monotonic Multi-Sender Signaling Games*. THEMA Working Papers 2022-08. THEMA (THÃ©orie Economique, ModÃ©lisation et Applications), UniversitÃ© de Cergy-Pontoise. DOI: None. URL: https://ideas.repec.org/p/ema/worpap/2022-08.html.