SCHOOL OF ECONOMICS

UNIVERSITY OF NOTTINGHAM

SOCIAL INFLUENCE IN ECONOMIC DECISION MAKING

ESSAYS ON COOPERATION AND MORAL JUDGEMENT

GEORG SATOR

September, 2025



Thesis submitted to the University of Nottingham
for the Degree of Doctor of Philosophy

Georg Sator: *Social Influence in Economic Decision Making: Essays on Cooperation and Moral Judgement*. Ph.D. in Economics. First supervisor Professor Simon Gächter. Second supervisor Professor Silvia Sonderegger. © Georg Sator, 2025.

To Nicol

”All that is simple is false, all that is not simple is useless.”

— Paul Valéry

Abstract

This thesis presents novel evidence of social influence from two controlled economic experiments. The first study examines peer effects in cooperative settings and shows that social closeness systematically moderates peer influence on contributions to a public good. By separating direct effects on choices from belief-based channels, we find that the impact of social closeness operates primarily through direct behavioural responses rather than through changed expectations. The second study investigates social influence on moral judgements and documents substantial conformity. Participants both genuinely adopt the moral views of others and strategically misrepresent their stated views when exposed to endogenous social reward and punishment mechanisms. Together, the results demonstrate that even minimal social cues can prompt substantive behavioural responses. We embed and interpret these findings within the broader literature on how the social environment shapes preferences, beliefs, and choices.

Acknowledgements

This thesis is about social influence, and thus, at its essence, about how what we do is shaped by others. Truly, the writing of this thesis itself has been profoundly shaped by many people who encouraged and supported me throughout my PhD journey. I am indebted to many people whom I would like to acknowledge.

First and foremost, to my supervisors, Simon Gächter and Silvia Sonderegger, for the advice they shared throughout these years. For the time and effort they invested, their guidance and support, for the discussions, and their genuine care beyond their role as supervisors. I never took this for granted.

I am especially indebted to Simon Gächter. His landmark insights into the puzzle of human cooperation first captivated me as an undergraduate at the University of Vienna. I was fortunate enough to have him as my supervisor for my master's at the University of Nottingham, and he continued to guide me through my PhD with sharp thinking and a warm heart as the mentor and academic advisor every student hopes for.

I would like to thank Fabian Dvorak, Urs Fischbacher, and Katrin Schmelz for our collaboration and for the many inspiring, good-humoured discussions that accompanied it.

I am grateful to Robin Cubitt for continuous feedback on my work since Year 1, and to him, together with Chris Starmer, for welcoming me into the CeDEx family. It has been a true privilege to be part of this inspiring research community. I have especially appreciated the intellectually sharp and well-intentioned environment that CeDEx cultivates across seminars, meetings, and hallway encounters.

I would also like to express my gratitude to the University of Nottingham School of Economics for awarding me a PhD scholarship and providing financial support to present my research in the UK and mainland Europe.

I owe special thanks to Armin Falk for inviting me to briq at the University of Bonn as a visiting PhD student twice, to Botond Kőszegi and Florian Zimmermann for extensive discussions of my research, to Devesh Rustagi and Seung-Keun Martinez for the many light-hearted yet illuminating conversations about research and the wider academic world, and to everyone who provided helpful comments at workshops, conferences, and seminars.

A special thanks to José Guinot Saporta for providing invaluable support in and around the lab, often well beyond regular working hours. And to Natalia Bas, Suzanne Robey and Samantha Stapleford-Allen for saving me from endless admin headaches.

I would like to thank my former colleagues in Vienna, Martin Kocher, Eryk Krysowski, Simon Martin, Wieland Müller, Karl Schlag, Florian Spitzer, Jean-Robert Tyran, and Alexander Wagner, and especially Ana Begoña Ania Martínez, for their feedback and advice, and for patiently answering my countless questions.

Many thanks to Philipp Gerlach for becoming a dear friend and intellectual sparring partner early on during my PhD. And to my colleague and flatmate, Richard Mills, for the endless conversations, laughter, food, and playlists we shared.

And finally, I am deeply grateful to my parents - for supporting me with unconditional love, wisdom, and tranquillity. To my brother Thomas, who I will always look up to. To Lisi and Frida, who are blessed with the gift of making me laugh just when I need it most. To Adriana and Laco, whose loving care both metaphorically and literally nourished me through all the highs and lows. And to my wife Nicol, for providing me with endless love, overwhelming support, lightness and humour - and for truly having my back.

Proofreading and editing of parts of the thesis, and German-to-English translations of experimental instructions and computer screens were done using ChatGPT (GPT-4 and GPT-5, OpenAI). All output was manually verified and revised.

Contents

Abstract **i**

Acknowledgements **ii**

List of Tables **vi**

List of Figures **vii**

General Introduction **1**

Chapter 1 Social Closeness and Peer Effects **10**

1.1 Introduction 11

1.2 Experiment 15

1.3 Hypotheses and Results 22

1.4 Discussion 47

1.5 Conclusion 48

Chapter 2 Social Influence on Moral Judgements **50**

2.1 Introduction 51

2.2 Design and Procedure 55

2.3 Results 64

2.4 Conclusion and Discussion 73

General Conclusion **74**

Bibliography **77**

Appendices **95**

A Chapter 1 **96**

A.1 Supplementary Figures 96

A.2 Supplementary Tables 100

A.3 Instructions 105

B Chapter 2 119

B.1 Supplementary Tables 119

B.2 Moral Problems 120

B.3 Classical Moral Problems 130

B.4 Instructions & Payout Info 132

B.5 Decision Screens 139

List of Tables

2.1 Linear probability models for evaluation behaviour 72

A.1 Structural Equation Model Estimates – Observed Participants . . 101

A.2 Decomposition of Effects by Treatment - Observed 102

A.3 Structural Equation Model Estimates – Observing Participants . 103

A.4 Decomposition of Effects by Treatment - Observers 104

B.1 Linear probability models for strength of preferences 119

List of Figures

1.1 Inter-group observation 18

1.2 Experimental timeline 20

1.3 Overview - observed participants 25

1.4 Estimated coefficients - observed participants 31

1.5 Overview - observing participants 37

1.6 Estimated coefficients - observing participants 43

2.1 Probability of intransitivity across cases and treatments 65

2.2 Intransitivity across treatments for each moral problem 69

A.1 IOS scores across treatments. 97

A.2 Similarity of cooperation preferences of friends vs. non-friends. 98

A.3 Differences in contribution choices of pairs of friends Phase 1 vs. Phase 2. 99

A.4 Decision screen for unconditional contribution choice in Part 1 . 110

A.5 Decision screen for conditional contribution choices in Part 1 . . 111

A.6 Illustration of observability as shown in instructions 116

A.7 IOS Task 118

B.1 Decision screen for decisions without social influence 139

B.2 Decision screen for decisions with social influence 140

B.3 Decision screen for reward decisions 141

B.4 Decision screen for punishment decisions 142

General Introduction

The Greek philosopher Aristotle famously noted that man, by nature, is a *zōon politikon*, a political animal (*Politics*, 1253a1–20). He highlights the uniqueness of the human species in its ability to communicate right and wrong through speech, affording them a communal understanding of what is just. A shared understanding of good and bad, in turn, he argues, provides the foundation for social and political life: "it is partnership in these things [i.e., perception of moral matters] that makes a household and a city-state" (*Politics*, 1253a15–20). (Arguably, a shared understanding of good and bad in politics remains work in progress in 2025.)

Since then (and even before Aristotle), a plethora of authors have devoted a great amount of thinking into the various ways in which individuals are influenced in their values, beliefs, and choices by other human beings, and more generally their social environment. Plato extensively discusses rhetorical persuasion (*Gorgias*, 455a), and Cicero mentions collective assertion as a source of wrong beliefs (*Tusculan Disputations*, III.2-3). Within the British Empiricists' tradition, John Locke stresses how reputation concerns lead humans to follow socially constructed notions of virtue (*An Essay Concerning Human Understanding*, II.xxviii.10-13), and David Hume argues that the human capability to empathise with others gives rise to moral sentiments (*A Treatise of Human Nature*, 3.3.1.7-10; SBN 575-78), which generalise to universal moral

rules through exchange with others (*An Enquiry concerning the Principles of Morals*, 5.42; SBN 228-29). The Genevan Enlightenment philosopher Jean-Jacques Rousseau used the term *amour-propre* to refer to relational self-love that depends on others' opinions, and which leads to social comparison, pride, shame, and status-seeking, potentially to the point where it breeds rivalry and harm (*Discours sur l'origine et les fondements de l'inégalité parmi les hommes* 1964/1755, esp. Note XV). With regard to how social influence shapes our beliefs, the German-Jewish political theorist Hannah Arendt points out the intersubjectivity of our perception of what is real, specifically and in economic terms, the critical role of others in the *affirmation* of our beliefs (1998/1958, p.50), while the French post-structuralist and historian Michel Foucault analyses how institutions and power relations shape *what* we regard as true (1980, p.131).

With regard to how our behaviour is shaped by social influence, many scholars in sociology have stressed the role of the internalisation of habits, customs and norms. Among them are Émile Durkheim (1938/1895, I.), Max Weber (1978/1956, I.2., p.25), Pierre Bourdieu (1977/1977, Ch.2), Talcott Parsons in his voluntaristic theory of action (1949; 1951), or Jürgen Habermas (1987, e.g., VI.1.B). Beyond voluntary or even automatic adherence to usual or appropriate behaviour, adherence to social rules has been argued to be *enforced* by means of external coercion, including exclusion from social groups and other forms of social sanctions (e.g., Émile Durkheim (1938/1895, I.) or Max Weber (1978/1956, I.6., p.34 & IX.6.E)). Granovetter (1985) rejects the claim that people primarily follow norms, as well as the notion of atomistic and rational agents. Instead, he emphasises that individuals are embedded in social networks which govern available information, trust, and normative judgements. Veblen (1922/1899, II-IV) analyses how the interplay of social comparison, reputation and status-seeking leads to emulation of the leisure and consumption styles of higher-status

groups (and thereby misdirects resources). And Kuran (1995) develops a theory according to which people's inclination to avoid social sanctions makes them act in line with socially acceptable norms, even if they go against their private values. In the aggregate, this creates the misleading impression that most people intrinsically support the norm, further solidifying it.

Also canonical works by social psychologists have discussed different forms of social influence. For instance, they examined conformity to group or majority opinions using the example of various kinds of visual perception (Sherif, 1935, 1936; Asch, 1951, 1956), obedience to authority against personal scruples (Milgram, 1963), and coined the term *social proof* for the tendency to treat others' common behaviour as a guide to optimal action in ambiguous situations (Cialdini, 1984). In a review article on compliance and conformity, Cialdini and Goldstein (2004) note that a person's susceptibility to social influence is motivated by the formation and preservation of accurate beliefs, social relationships, and a favourable self-concept.

Evidently, throughout the history of philosophy and social science, the question of how people influence each other in their thoughts and deeds has been of central interest. Unsurprisingly, economics is no exception.

Microeconomics, at its core, examines the production, allocation, and consumption of scarce goods. Through the institution of markets, individual entities meet and exchange goods, services, and labour. A firm earns profit only if consumers are willing to buy its output, a consumer can purchase only what others choose to produce, and a worker's consumption possibilities are bounded by the wage that an employer pays in exchange for labour.

In this sense, a basic form of social influence is intrinsic to economic life and thus at the very heart of what economists study.

But even beyond mere exchange, the history of economic thought is full of instances of social influence between economic agents. One channel more generally involves decision problems in which agents mechanically influence each other because of strategic interdependence: the actions of one agent affect the payoff of another by nature of the decision problem. For instance, early analyses of oligopolistic competition fall into this category (Cournot, 1897/1838; Bertrand, 1883[1]; Edgeworth, 1925/1897).[2] Bargaining, e.g., over how to split a pie by making alternating offers which the other can either accept or reject (Rubinstein, 1982), or classic textbook games such as Prisoner's Dilemma, the Dating Game or Stag Hunt are further canonical examples.

Another channel operates through information transmission: observing someone's choices may allow another individual to better understand how to achieve optimal outcomes for themselves. For example, when many individuals sequentially face similar decisions and only have noisy private information about which option would be best, the choices of first-movers can dominate later agents' private information, leading to information cascades and potentially suboptimal outcomes for those who choose later (Banerjee, 1992; Bikhchandani et al., 1992; Welch, 1992; Bikhchandani et al., 1998; Anderson and Holt, 1997). Even prior to these insights, Hayek (1945) showed how influence can be mediated by prices. He argued that dispersed private information is aggregated in prices and allows economic agents to make inferences about relative scarcities, and thus about costs and demand.[3]

Many important instances of economic agents influencing each other sit at the

[1] For a discussion of the analyses by Cournot and Bertrand, and an appended English translation of Bertrand's text, see de Bornier (1992).

[2] The modern axiomatic theory of strategic situations was pioneered by von Neumann and Morgenstern (1944), and extended by Nash Jr. who introduced the noncooperative framework of self-enforcing equilibria (Nash Jr, 1951) and proved their existence for finite games (Nash Jr, 1950). See Myerson (1999) for a historical perspective on Nash Jr.'s contributions to game theory.

[3] Though later work added a caveat to this, arguing that nobody would pay to acquire information if it was *fully* contained in prices (Grossman and Stiglitz, 1980).

intersection of these two channels. For example, agents may choose costly actions to influence subsequent choices of another agent in a way that is payoff-relevant for both. Spence (1973) considers a setting in which acquiring education, without increasing productivity, is less costly for high-ability workers. Doing so hence credibly reveals to employers that a worker is a high-ability type, thus inducing the employer to pay a higher wage. Or, if a firm decides to enter a market, incumbents may cut the new entrant's profit by setting prices aggressively low, thereby building a reputation that deters potential future entrants (Kreps and Wilson, 1982). Crawford and Sobel (1982) show that if preferences are sufficiently aligned, also costless messages that do *not* change anyone's payoffs directly can be sufficiently informative (as opposed to merely uninformative cheap talk) to steer the action of the receiver of the message in a way that increases both the sender's and the receiver's payoffs. Already prior to that, Schelling (2015/1960, especially Part II, Chapter 5) argued that constraining one's own future options through credible commitment, threats, or promises can induce an opponent to best respond as intended. An agent may, for instance, sign a contingent contract that specifies how they will respond in a given state and that imposes a penalty for deviation. Knowing that not following through is costly for the agent, the opponent anticipates this (new) consequence and may now rationally shy away from a previously optimal choice.

Contrary to social influence based on information transmission or strategic interdependencies, the evidence of social influence presented in this thesis is *in*consistent with decision makers caring about their individual wealth only. It thus connects to a large literature which remains inside the rational choice and equilibrium paradigm, but incorporates so-called *non-standard preferences*. Instead of operating through strategic interdependence or information transmission alone, social influence then arises because agents' preferences themselves depend on social context.

The list of preference enrichments that incorporate social context to account for behavioural patterns that the purely material, self-interest benchmark leaves unexplained is long. This diversity of models reflects the fact that different types of decision problems and social cues activate distinct psychological motives, so no single specification fits all settings.

With respect to social influence, foundational mechanisms by which one agent's actions affect others include altruism (Becker, 1974), altruism and associated status (Bénabou and Tirole, 2006), warm-glow from the act of giving (Andreoni, 1990), intrinsically motivated conformity to social norms (Elster, 1989) or associated status gains (Bernheim, 1994; Akerlof, 1980, 1982), preferences for reciprocating kind or unkind intentions (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004), perceived fairness of the outcome (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), the anticipation of punishment by people who are willing to incur costs to sanction norm violations (Fehr and Gächter, 2000, 2002; Boyd et al., 2003), or a preference for aligning actions with socially shaped identities (Akerlof and Kranton, 2000). The economic implications of each of those are manifold, playing out in a variety of contexts, including charitable giving, wage levels, unemployment rates, cooperation, gender roles, and many other aspects of social life.[4]

Chapter 1 and Chapter 2 both present original evidence of social influence stemming from controlled experiments. While largely agnostic about the underlying psychology as such, both chapters document the existence of empirical patterns that cannot be readily explained by any of these theoretical accounts. This is the main contribution of this thesis.

In Chapter 1, we examine social influence in the context of cooperation, focusing on the causal role of subjective closeness between peers.

[4]For a review of the social preferences literature with a special focus on economic consequences, see Fehr and Charness (2025).

Cooperation is unanimously considered a central pillar of the success of the human species. A cooperative choice is one that forgoes selfish benefits for the greater good of others. The work that has been done on understanding cooperation, how it emerges, how it can be sustained, and when it collapses is vast and spans several decades of work across various disciplines (see, for example, Fehr and Gächter (2002); Fehr and Fischbacher (2003); Bowles and Gintis (2003); Gächter and Herrmann (2009); Fischbacher and Gächter (2010); Curry et al. (2019); Efferson et al. (2024)).

The class of social influence we study here is sometimes referred to as *peer effects*. Their effect on behaviour arises from the presence of others who, under assumptions typically made in economic analysis, are not even part of the decision environment. In principle, peer effects may emerge whenever choices are observable, and whatever the context.

As will be discussed in the introductory section of Chapter 1, peer effects have been shown to significantly shape human behaviour in many important settings. Moreover, every peer-to-peer interaction is necessarily characterised by some degree of social closeness, which, however, varies from one pair of peers to the next. This warrants a systematic examination of the role of social closeness in peer effects.

Chapter 1 presents a controlled laboratory experiment that employs a novel procedure which allows us to study this question in the context of cooperation. It considers both types of peer effects: those arising when observing a peer, and those caused by being observed by a peer. The experiment is constructed to ensure payoff externalities, social-preference spillovers, and learning about the game are ruled out as sources of peer effects.

A key methodological innovation in our experiment is that we vary social closeness between peers in three steps, from mutual anonymity to friendship, while

holding constant all other parameters of the decision environment. It further allows us to control for potential *ex ante* existing similarities in cooperation types and expectations of cooperativeness of others by using the ABC method by Fischbacher and Gächter (2010).

In our analyses, we distinguish between direct and indirect effects transmitted through beliefs. With regard to observed participants, we find that peer effects decline with social closeness, and there is no evidence for additional effects on beliefs. For observing participants, we find that responsiveness to observed choices increases with social closeness. As for observed participants, we do not find social closeness to moderate belief-based indirect effects.

Chapter 2 reports on a study situated in the domain of moral behaviour. Morality, a cornerstone of civilisation, plays a central role in human decision making, including economic decision making, in shaping our political views, and in cooperation.[5] The simple fact that much of our moral deliberation and action takes place in the presence of others raises the question of whether or how they are shaped by social influence.

The experiment presented in Chapter 2 addresses this question. In an online experiment, we elicited participants' moral judgements on a series of dilemmas while systematically varying the forms of social influence to which they were exposed. A central aim of our study was to distinguish genuine changes in participants' judgements, prompted by the opinions of others, from strategic responses driven by expectations of social reward or punishment. To achieve this, participants were assigned to one of three treatments. In one, they learned the moral judgements of others before stating their own. In the other two, they were additionally informed that their judgements would be evaluated against those of others, with the possibility of receiving extra money or incurring deductions. A

[5] See, e.g., Haidt (2007); Graham et al. (2009); Enke et al. (2023); Gintis et al. (2008); Enke (2019); Curry et al. (2019).

control condition established a benchmark for changes in moral judgements net of any form of social influence.

The results reveal significant effects of social influence. Specifically, participants' judgements exhibit strong conformity when informed about others' opinions. This tendency is further amplified by strategic incentives, resulting in even stronger conformity. Notably, these patterns do not merely reflect average effects across the set of moral problems, but also emerge consistently at the level of individual dilemmas.

Finally, this thesis concludes with a summary of the key insights on social influence drawn from Chapters 1 and 2, considered both individually and as a whole.

Chapter 1

Social Closeness and Peer Effects

Abstract

This paper studies peer effects in a cooperation experiment under exogenously varied degrees of social closeness. Two principal forms of peer effects are considered - those induced by observing, and those induced by being observed by a peer. Our results show that social closeness plays an important role in both. While the effect of being observed *decreases* with social closeness, the influence on observing peers *increases*. A key contribution of our study is the combination of a novel experimental design with structural equation modelling, which allows us to disentangle preference-based from belief-based channels. We interpret our findings through the lens of reputation and social learning mechanisms.

This chapter is based on joint work with Simon Gächter and Silvia Sonderegger. We thank Johannes Abeler, Robin Cubitt, Urs Fischbacher, Philipp Gerlach, Seung-Keun Martinez, Devesh Rustagi, Peter Schwardmann, Florian Zimmermann, Ro'i Zultan, as well as seminar participants in Konstanz, MPI Berlin, Nottingham, University of Vienna, and WU Vienna, and participants at the ESA Meeting Osaka 2022, AFEE Dijon 2021, Tiber Symposium Tilburg 2021, RExCon21, IAREP-SABE 2021, SABE 2020, BEEC Bogotá 2020, NoBeC University of Pennsylvania 2019, ESA Meeting Dijon 2019, and ECEM Conference Essex 2019 for helpful comments.

1.1 Introduction

Cooperation is a key feature of humanity. It enables us to live in and benefit from social groups and thus constitutes the foundation of civilisation and human co-existence. Despite a large and enlightening literature, the empirical fact of cooperation in one-shot anonymous interactions remains a challenge for the social sciences. In this paper, we study how social closeness shapes peer effects in cooperation problems.

Peer effects exist if an agent's actions are influenced by a comparison agent even in the absence of material spillovers and instrumental learning about payoff-maximising play. According to this definition, peer effects persist regardless of whether direct social preference links between peers are possible, and may operate through channels involving social identity, reputation, or shaping beliefs about norms. Peer effects are an interesting potential determinant of cooperative behaviour because they can be caused by (seemingly) entirely uninvolved parties.

Peer effects have been shown to play an important role in diverse settings, including education (Sacerdote, 2001; Bursztyn and Jensen, 2015; Bursztyn et al., 2019a), the workplace (Mas and Moretti, 2009; Falk and Ichino, 2006), criminal behaviour (Bayer et al., 2009), charitable giving (DellaVigna et al., 2012), participation in social programs (Dahl et al., 2014), voter turnout (Gerber et al., 2008; Funk, 2010), financial decision making (Bursztyn et al., 2014), energy consumption (Allcott, 2011; Ferraro and Price, 2013), adoption of environmentally friendly technology (Bollinger and Gillingham, 2012), reciprocal gift exchange (Thöni and Gächter, 2015). Common to these findings is that peer effects manifest in one of two principal forms which are driven by distinct psychological mechanisms. One, caused by being observed by a peer, has been associated with image or reputation concerns (e.g. Fehr (2004)). The other, caused by ob-

serving a peer, can prompt conformity to choices or perceived norms (e.g. Thöni and Gächter (2015); Gächter et al. (2013)), and is driven by how an individual's choices or outcomes compare to those of others. In this study we consider both types.

Importantly, peer effects occur in social relationships of varying degrees. Ranging from mutual anonymity to close friendships, social closeness is an inevitable feature of any real-world peer-to-peer environment. And since, for example, people tend to care more about their reputation vis-à-vis a friend than a complete stranger, it plausibly determines peer effects (Gächter et al., 2023). Defined as unilateral and subjective perception of how close someone feels to another person, social closeness is entirely up to the individual and, as such, neither needs to be linked to anything objective nor reciprocated. It is shaped by quality and frequency of interactions and regarded as a key characteristic of interpersonal relations in the psychology literature (Berscheid et al., 1989).

However, identifying the causal role of social closeness in peer effects remains elusive. It is inherently difficult to measure because of its subjective nature. And because people typically *choose* with whom they interact, the roles they assume, and the environments in which these interactions occur, social closeness is often endogenous. Moreover, in repeated interactions, potential initial influence of social closeness can be obscured by shared norms and conventions that evolve jointly over time. Methodologically, laboratory experiments usually impose complete anonymity, while field environments seldom offer exogenous manipulation. Thus, they often lack the exogenous variation in closeness that would allow its effect to be cleanly measured. As a result, despite being inevitable in many social interactions, the literature offers limited insights into how social closeness shapes peer effects.

This paper seeks to fill this gap by means of a novel experimental design. We

choose a controlled lab experiment for multiple reasons.

First, a novel procedure allows us to overcome endogeneity issues which are otherwise difficult to account for. The core design feature consists of exogenous variation in how interacting peers relate to each other, holding other elements of the decision environment constant. We verify the psychological meaning of this procedure using the oneness measure by Aron et al. (1992). A comprehensive evaluation by Gächter et al. (2015) attests that the oneness measure is psychologically meaningful and highly reliable as an experimental tool to measure subjectively perceived closeness of relationships.

Second, the controlled lab conditions allow us to shed light on the mechanisms that underlie peer influence, because they permit the isolation of behavioural responses caused by the interaction with a peer in direct effects on choices, from indirect effects transmitted through shifts in expectations about the choices of others. As we will discuss in detail below, this is crucial when studying peer effects in strategic settings.

Third, drawing on the ABC method developed by Fischbacher and Gächter (2010) and Gächter et al. (2017) we can construct a benchmark of how each individual participant *should* behave *in the absence of any peer influence*. This approach, which measures preferences and beliefs, has consistently been shown to generate remarkably accurate predictions by using an individual's behavioural type and pairing it with their expectation about the choices of others. We use it to understand whether the observed levels of cooperativeness already existed *ex ante* or were caused by treatment. This is of particular importance for studies involving friends. Due to homophily, friends may be more similar than two random strangers, including preferences relevant in contexts of cooperation. The ABC method helps us to separate pre-existing similarities in preferences from potential similarities in choices caused by, e.g., imitation.

Lastly, previous literature suggests that peer effects found in the lab are not just an artefact of its sterile conditions. The estimated effect sizes found in the lab have been shown to closely parallel findings in the field (Herbst and Mas, 2015).

By shedding light on the role of social closeness in peer effects, we aim to reproduce a broader and more accurate representation of diverse peer-to-peer environments. This has important practical implications, as in many real-world settings where interventions seek to promote desirable behavioural change, peer effects often constitute a key mechanism of influence. Examples include retirement plan decisions (Duflo and Saez, 2003; Beshears et al., 2015), water (Ferraro et al., 2011; Ferraro and Price, 2013; Bernedo et al., 2014) and energy consumption (Allcott, 2011; Ayres et al., 2013), the adoption of new health products such as mosquito nets (Dupas, 2014), educational investment (Bursztyn and Jensen, 2015), vaccination take-up decisions (Sato and Takasaki, 2019), female genital cutting (Grose et al., 2019) or participation in programs aiming to prevent blackouts (Yoeli et al., 2013). Successfully leveraging peer effects requires a concise understanding of how they change relative to specific prevailing circumstances, including social closeness between peers.

The results of our experiment constitute causal evidence that social closeness determines peer effects both for observed and observing participants. Thus, our study informs two common queries in the literature. First, we replicate the result that people become more prosocial when their choices are observed (Bradley et al., 2018) and show that this pattern is sensitive to variations in social closeness: cooperativeness significantly increases among participants who are observed by strangers, but this effect gradually decreases and even vanishes as social closeness increases. Second, the results of our experiment speak to the literature on conformity and imitation, where people are often found to make choices similar to those of their peers (Falk and Ichino, 2006). We find that the way observed choices determine one's own choices significantly differs between

low and high social closeness.

The remainder of the chapter is organised as follows: Section 1.2 describes the experimental procedure and design, Section 1.3 shows the results, Section 1.4 provides a short discussion, and Section 1.5 concludes.

1.2 Experiment

Here we provide a complete picture of the experiment and how it was run. Running the experiment required a relatively unusual recruitment and pre-experimental procedure. We therefore first explain these procedures, before we turn to the details of the design.

1.2.1 Procedure

We recruited participants in pairs of friends. This was done by sending out invitation emails using the recruitment platform ORSEE (Greiner, 2015). It included a passage stating that participation was only granted upon meeting the following two requirements: First, participants had to indicate the name, first language and field of study of a friend of theirs (who may or may not be a member of the lab's subject pool). Second, this friend must come to participate in the same session. IDs of initially invited participants and friends were checked upon arrival at the lab to verify that the announced friend was the one actually present. All other steps of the recruitment procedure followed regular protocols.

Once IDs had been verified, each pair of friends drew a card which indicated both of their respective seat numbers. Beyond the physical partitioning of the lab, seat numbers were pre-assigned on each card to ensure that the two friends were positioned sufficiently far apart to prevent any possibility of communica-

tion. The use of mobile phones was strictly forbidden, and participants were informed that any violation of this rule would lead to exclusion from the experiment. At their seats, participants found a sealed envelope, which they were instructed not to open until further notice. Only once all participants were seated were they asked to open the envelope and follow the instructions they found in it. The instructions directed half of the participants to remain at their original seats, while the other half were asked to move to the seat corresponding to the number they had been assigned. Through this procedure, each participant was matched with one other previously unknown participant. The possibility of friends meeting each other was excluded by design. Furthermore, pairs were formed such that, due to the distance between their seats, no communication was possible once they had returned to their seats and the actual experiment began. They then had two minutes to talk about anything they wished. After the two minutes had passed they were asked to go back to their seats. Only then were further instructions handed out and the actual experiment began. Thus, at the time they encountered the other person they had no information about what the remainder of the experiment was going to be about.

To ensure that results would not be confounded by differences in recruitment or the procedure participants had been exposed to, both the recruitment of pairs of friends and the introduction to a stranger were held constant across all treatments, even when this was not strictly required by the treatment implemented in a particular session.

276 subjects participated in the experiment[1]. A typical session lasted between 1.5 and 2 hours. Immediately after a session had ended, participants were paid

[1]Four of the 276 participants had to be excluded from the analysis due to a technical problem with one of the computers. One participant switched off (and back on) his computer before the experiment started, which caused a mismatch of in total four participants. All other participants were unaffected. The incident happened in the course of participants entering the lab and taking their seats. It went unnoticed by lab personnel and only became apparent after the experiment had ended, because seat numbers and participant numbers on the payment file did not match.

the money they had earned. €4854.26 were paid out to them in total, and average earnings amounted to €17.59. Participants were mostly students, on average 24 years old, and 51% of them were female. The experiment was programmed in z-Tree (Fischbacher, 2007) and conducted at the University of Vienna.

1.2.2 Design

We chose to study peer effects in the context of cooperation and therefore employed the canonical public goods game as our experimental framework. Following the respective contribution decisions, we elicited participants' beliefs about the average contributions of their group members in an incentive-compatible manner.[2] participants were randomly assigned to groups, subject to the restriction that neither friends nor participants introduced to one another during the experiment's pre-phase could be placed in the same group. This restriction, which was known to participants, was intended to prevent confounding contribution choices with potential interaction effects between social preferences and social closeness to their group members.

The experiment consisted of two phases. Phase 1 was the same for all participants irrespective of treatment. Its sole purpose was to elicit participants' *ex ante* contribution preferences by means of a strategy method public good game. Phase 2 constituted the main part of the experiment. It used a direct-response version of the same public good game and supplemented it with controlled peer-to-peer interaction (details below). Across treatments, the social closeness between peers was varied exogenously in a between-subjects design. Participants only learned what Phase 2 was going to be after Phase 1 had ended. They further knew that Phase 2 would be the final phase of the experiment. Moreover, any

[2]Participants received three experimental currency units for a correct guess, two units if their guess was off by one, and one unit if it was off by two.

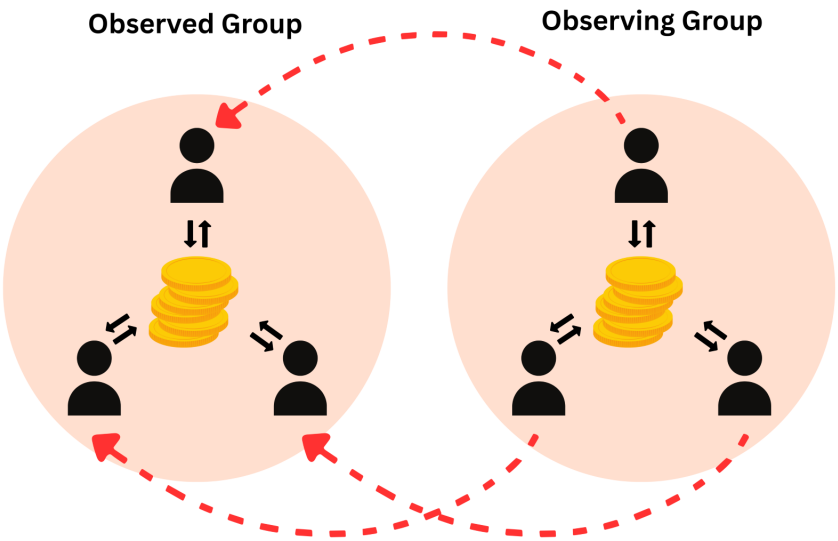


Figure 1.1: Inter-group observation. Two payoff-independent public good groups formed a pair consisting of one observed and one observing group. Each member of an observing group exclusively observed the contribution choice of exactly one member of an observed group. Observed group members knew they were going to be observed by the time they made their contribution choice.

outcome of Phase 1, such as earnings or other players' contributions, were not disclosed until the very end of the entire experiment.

Parameterisation and monetary incentives were identical in Phases 1 and 2. The payoff function in each phase was given by $\pi_i = \alpha \sum_{j=1}^{N} c_j + e - c_i$. Each group consisted of $N = 3$ players endowed with $e = 20$ experimental currency units each (henceforth *ECU*). The marginal per capita return per ECU invested in the public good was given by $\alpha = 0.5$. At the end of the session participants were paid out according to how many ECU they had earned throughout the experiment. The conversion rate was €0.30 per ECU.

To set the stage for (potential) peer effects we implemented peer-to-peer interaction in the form of observability of choices between players in Phase 2. Figure 1.1 illustrates the basic concept.

We randomly divided public good groups into equal numbers of so-called *ob-*

serving groups and *observed groups*. Decisions were made sequentially across these two group types. First, all members of observed groups made their contributions to their respective public good. Subsequently, each member of an observing group was shown the contribution choice of *one* member from an observed group. Only then were members of observing groups asked to make their own contribution decisions. Observed and observing players were matched one-to-one: each member of an observed group was observed by exactly one observing player, and each member of an observing group observed one player. Consequently, all observations were exclusively made *across* payoff-independent public good groups. This design feature is crucial, as the exclusion of payoff spillovers between observed and observing peers is necessary to cleanly isolate peer effects as defined above - *social* spillover effects in the absence of social preference links. This entire procedure was made common knowledge before any player, observed or observing, had made a choice. Thus, by the time observed players made their contribution decisions, they already knew that their choice was going to be observed by a member of an observing group.

The treatment variation was implemented in a between-subjects design and manipulated the degree of social closeness between observed and observing participants. In treatment *Friend*, observations took place between the respective friends who had come to the lab together. In *Stranger*, observations were made between participants who had briefly met and talked to each other during the pre-experimental phase. In *Anonymous*, observations occurred between neither of them but between two mutually anonymous participants randomly matched within the session. The treatment-specific matching procedure was made common knowledge to all participants before any contribution decisions had been made - both by observed and observing participants. In the control condition *No Observation* all participants made their contribution decisions without observing others or being observed themselves.

To verify the assumption that our treatment variation induces meaningful variation in perceived social closeness, we asked participants to rate how close they felt to the observed/observing player using the IOS test by Aron et al. (1992). This was done immediately after Phase 2 had ended. The results confirm that the treatment variation has psychological relevance. Average test scores by treatment are 2.42 in *Anonymous*, 2.83 in *Stranger* and 4.79 in *Friend*[3]. The upward trend is statistically significant at the 1% level based on the Jonckheere–Terpstra test.

Finally, the experiment ended with the social value orientation task by Murphy et al. (2011)[4] and a post-experimental questionnaire. The timeline of the experiment is illustrated by Figure 1.2.

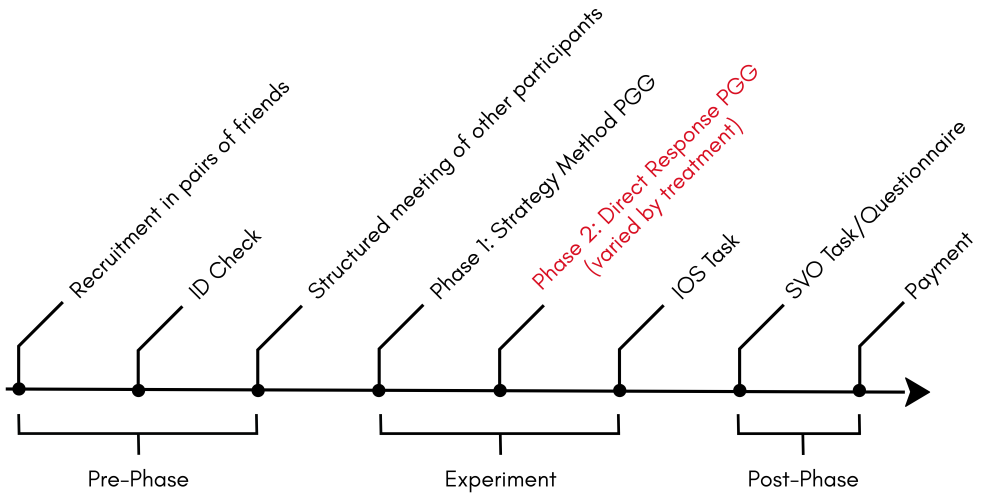


Figure 1.2: Experimental timeline. Treatment variation (between-subjects) was implemented in Phase 2 (*highlighted in red*): peer observability with exogenous variation in social closeness. In the control condition, no observability was introduced.

[3]See Figure A.1 in Appendix A.1 for histograms of IOS scores across treatments.
[4]We used the z-Tree implementation provided by Crosetto et al. (2012).

1.2.3 Discussion

To address our research question, we aim to isolate the role of social closeness in the non-strategic primary peer effect and indirect effects running through beliefs. Identification is achieved through random assignment to treatment while (i) controlling for pre-existing cooperation preferences and (ii) separating direct from indirect channels running from treatment assignment to contribution choices.

We designed the experiment such that, if social closeness matters for peer effects and peer effects matter for cooperation, then cooperativeness should differ systematically across treatments. Importantly, such systematic differences can be caused by psychological mechanisms that, arguably, lie outside our definition of peer effects. For example, participants in observed groups may reciprocally raise their contributions if they believe their group members make high contributions in an attempt to signal prosociality - even if they themselves are unaffected by their own observability. In this case, the behavioural response to treatment is caused by *others'* peer interactions instead of their own. This contrasts with observing participants who update their expectation of their group members' average choice based on their own peer's contribution, and subsequently alter their own contribution accordingly. As in the previous example, the effect of treatment on contribution choices is *indirect*, transmitted through beliefs, rather than direct. However, contrary to the previous example, it stems from the individual's own interaction with their peer, and thus qualifies as a possible channel through which peer effects may operate, instead of a confounding mechanism.

Eliciting beliefs further allows us to control for potential *ex ante* existing differences (or similarities) in cooperation preferences. This is important, as previous literature shows that cooperative behaviour is significantly shaped by attitudes towards cooperation and beliefs about the cooperativeness of others (Fis-

chbacher and Gächter, 2010). By pairing the cooperation profiles elicited in Phase 1 with the beliefs elicited in Phase 2, we can construct a benchmark for participants' contribution in the absence of (direct) peer influence and hence isolate the role of peers from other determinants of contribution choices, including effects transmitted through beliefs. This is especially helpful when evaluating whether similarities in contribution choices of friends stem from observation or from pre-existing alignment in cooperation preferences.

Lastly, the one-shot characteristic of Phase 2 allows us to observe participants' genuine cooperativeness (and how it is affected by peers) by avoiding the strategic complexity of repeated interactions. In repeated interactions positive contributions do not necessarily reflect genuine cooperativeness, but potentially an attempt to induce reciprocal contributions by group members in subsequent periods.

1.3 Hypotheses and Results

Peer effects have two dimensions - being observed by a peer and observing a peer. They largely connect to different strands of literature and evoke different psychology. Accordingly, this section is divided into two parts: one analysing the behavioural responses of observed participants, and the other focusing on those of observing participants. Our main outcome variable is cooperativeness, measured in contributions to a public good. In both parts we will first provide an overview of the contribution behaviour across treatments and control, and subsequently isolate the peer effect as a function of social closeness from confounding factors – most importantly from indirect effects caused by changed beliefs and thereby induced reciprocal responses.

1.3.1 Being observed

Hypotheses

The defining feature of cooperation problems is the fundamental trade-off between individual self-interest and collective welfare. Under the assumption that participants do not want to appear greedy, we expect higher contributions of observed participants compared to unobserved participants in the control condition *No Observation*. This aligns with a large body of literature showing that being observed triggers reputational psychology and increases individuals' propensity to cooperate (Rege and Telle, 2004; Bateson et al., 2006; Rand and Nowak, 2013; Grimalda et al., 2016).

The effect of social closeness on the contribution levels of observed participants in *Anonymous*, *Stranger*, and *Friend* is theoretically ambiguous *ex ante*. This ambiguity arises from the fact that social closeness can affect reputation concerns in two opposing ways, both of which are rooted in the strong association between social closeness and frequent, diverse, and meaningful interactions (Gächter et al., 2015).

Under the logic of indirect reciprocity, incentives to appear cooperative increase if future interactions are more likely. Hence, people have a stronger incentive to build (or maintain) a prosocial reputation vis-à-vis friends than vis-à-vis strangers whom they will likely not meet again.

On the other hand, individuals may have already established a stable reputation vis-à-vis close peers - such as friends - through repeated past interactions. As a result, the marginal value of sending an additional signal of cooperativeness is likely to be low. This reduces the reputational cost of selfish behaviour when being observed by a close friend, thereby lowering the opportunity cost of con-

tributing less. Thus, individuals who are observed by a friend may be more inclined to prioritise material gains compared to individuals who are observed by a previously unknown stranger. In this case, social closeness would exert a negative effect on contribution behaviour.

Which of the two effects prevails is an empirical question - and one that motivates our experiment. If the former does, then contributions of observed participants will on average be highest in *Friend*, followed by *Stranger*, and lastly *Anonymous*. If the latter effect dominates, then the reverse order should hold.

Results

Overview of data. We begin the presentation of the results by relating contributions to beliefs and *ex ante* propensities for cooperation. A corresponding visualisation of the data is provided by Figure 1.3.

Previous literature (e.g. Fischbacher and Gächter (2010); Gächter et al. (2017, 2022)) has established that contribution behaviour is largely determined by individuals' cooperation types and their beliefs regarding the contributions of others. Broadly speaking, our findings are consistent with this pattern. A majority of participants classify as conditional cooperators[5]. Consistent with that, we see a clear positive correlation of beliefs and contributions. Free-rider types, on the other hand, are predicted to contribute (close to) zero regardless of their beliefs. This concurs with the accumulation of dots at *contribution* = 0 (and slightly above) in *No Observation*, *Stranger*, and *Friend*. Although treatment allocation was random and a χ^2 test does not reject the null of equal type distribution across treatments ($p = 0.42$), there appears to be a slight undersampling of free rider types in *Anonymous*. Thus, the flat line at *contribution* = 0 is absent in

[5]In our classification of participants in contribution types we follow Fischbacher et al. (2001).

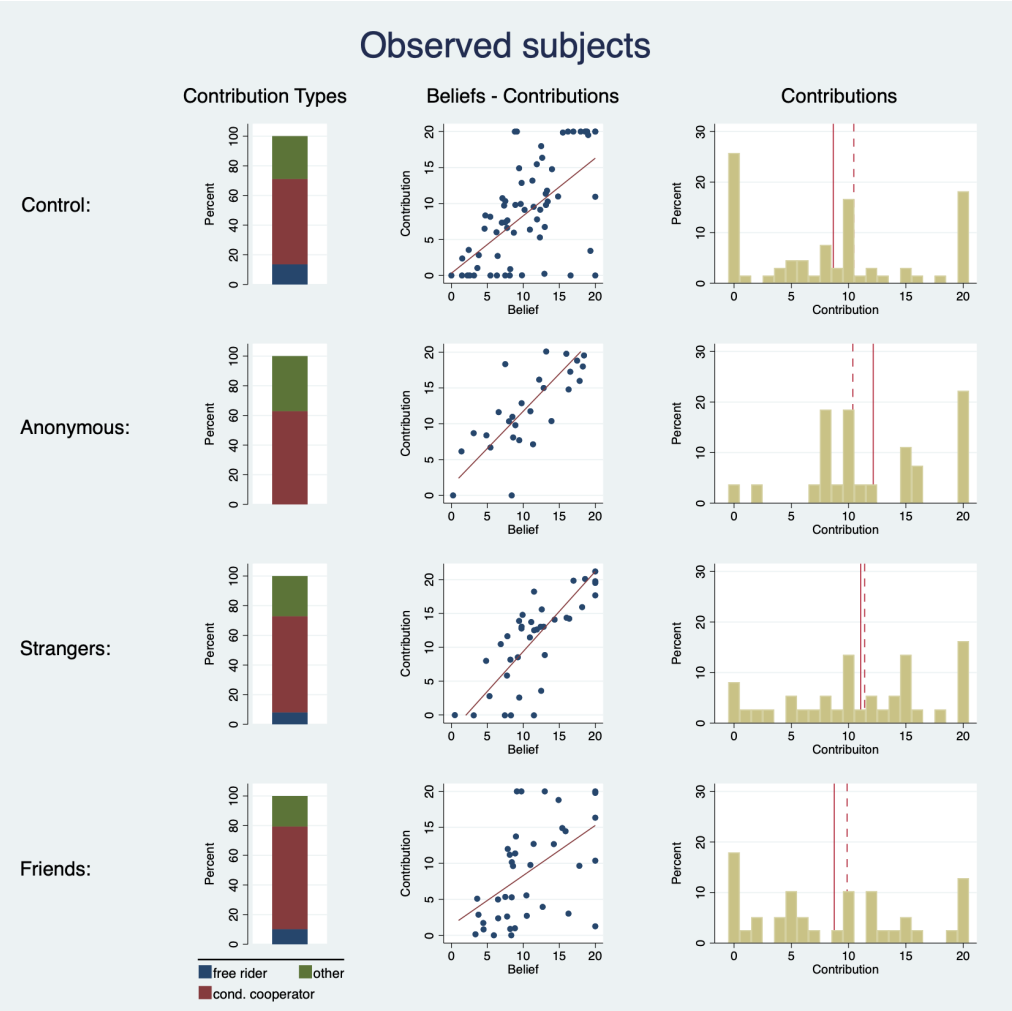


Figure 1.3: Overview - observed subjects. Left: distribution of contribution types based on strategy profile from Phase 1. Centre: Scatterplot of individual contributions and beliefs about group members' contributions in Phase 2, with fitted regression line. Right: Histogram of contributions in Phase 2, with mean contributions (solid line) and mean beliefs about group members' average contribution (dashed line).

the corresponding contribution-belief scatter of Figure 1.3. We control for this in the subsequent analyses.

The right column shows the resulting distribution of contribution choices, which differs noticeably across treatments. This supports the well-established finding in the literature that individuals tend to be more cooperative when observed. Relative to *No Observation*, participants contributed 40% more in *Anonymous*. This effect, however, attenuates as social closeness to the observer increases. In *Stranger*, the increase was 27.5%, and in *Friend* the effect disappeared entirely, with contributions virtually identical to those in *No Observation* (∼0.01% higher). This downward trend in mean contributions is statistically supported by non-parametric tests at the 1% level.[6]

What does this imply about the role of social closeness in peer effects? At least *prima facie*, the results suggest that social closeness mitigates reputation concerns, thereby lowering contributions to levels comparable to those of unobserved participants in *No Observation*. Put differently, the magnitude of peer effects appears to be inversely related to the degree of social closeness to the observer. One may infer that observed participants attach little value to signalling cooperativeness to their friends, vis-à-vis whom they have an established reputation already, while they seem to be willing to forgo the monetary benefits of selfishness to build a prosocial reputation vis-à-vis previously unknown observers.

However, this argument alone is not conclusive. As expected, beliefs seem to play a significant role in shaping contribution choices. Thus, anything that affects beliefs, including the treatment itself, may indirectly affect contribution choices. Since we define peer effects as any changes in an individual's behaviour that occur due to them interacting with their peer, it is necessary to control for

[6]Mean contributions *No Observation*: 8.67 < *Friend*: 8.74 < *Stranger*: 11.05 < *Anonymous*: 12.15 ($p < 0.01$, Jonckheere–Terpstra).

the possibility that participants adapt their contributions as a response to expecting their group members to respond to being observed. To do so, we estimate a structural model consisting of three simultaneous equations. It allows treatment to affect contribution choices not only directly, but also indirectly via beliefs, and enables us to conveniently quantify both of these mechanisms. Thus, if the previously reported differences in contribution choices across treatments are *not* caused by direct, generic peer effects but instead reflect expectations of how others react to treatment, this should become apparent in the model.

Estimation strategy. The structural equations are given by Equation 1.1:

$$\text{Belief}_i = \beta_0 + \beta_1 A_i + \beta_2 S_i + \beta_3 F_i + \varepsilon_i^{(a)} \tag{1.1a}$$

$$\text{ABCpred}_i = \gamma_0 + \gamma_1 \text{Belief}_i + \varepsilon_i^{(b)} \tag{1.1b}$$

$$\text{Contribution}_i = \delta_0 + \delta_1 A_i + \delta_2 S_i + \delta_3 F_i + \delta_4 \text{ABCpred}_i + \delta_5 \text{Belief}_i + \varepsilon_i^{(c)} \tag{1.1c}$$

Contributions are the main outcome variable of our model. As indicated by Equation 1.1c, contributions are jointly determined by the dummy variables A_i, S_i, and F_i, representing allocation to the treatments *Anonymous*, *Stranger*, and *Friend*, the ABC prediction, and beliefs about the contribution choices of others.

The coefficients of the treatment dummies, δ_1, δ_2 and δ_3, capture the direct effects of treatment allocation on contribution choices, formally given by

$$\text{Direct Effect} = \left.\frac{\partial\, \text{contribution}_i}{\partial\, \text{treatment}_i}\right|_{\text{belief}_i,\ \text{ABCpred}_i}.$$

The ABC prediction accounts for individual variation in cooperative dispositions. It is formally given by $\hat{c}_i = a_i(b_i)$, with b_i denoting beliefs, and $a_i(.)$ a

function which maps these beliefs to contributions according to an individual's attitude towards cooperation as elicited in Phase 1.[7] Lastly, because previous studies using the ABC method have consistently found beliefs to influence contribution decisions beyond what is captured by the ABC prediction alone, we include them as a separate covariate, despite already being embedded in the ABC prediction.[8]

Equation 1.1b captures the link between beliefs and the ABC prediction. Since, by construction, the ABC prediction is determined by beliefs, omitting Equation 1.1b in our model and treating the ABC prediction as exogenous would downward bias our estimate of indirect effects transmitted through beliefs, only capturing the influence beliefs have *beyond* the ABC prediction.

Finally, Equation 1.1a represents how treatment allocation shapes expectations about group members, feeding back into contribution decisions both directly and via the ABC prediction - together forming the indirect effect of treatment allocation on contribution choices.

With $\text{Treatment}_i \in \{A_i, S_i, F_i\}$, an indicator variable for the three treatments involving observability, the indirect effect of treatment on contributions via beliefs is formally given by

$$\text{Indirect Effect}_{\text{belief}} = \frac{\partial\, \text{belief}_i}{\partial\, \text{treatment}_i} \times \frac{\partial\, \text{contribution}_i}{\partial\, \text{belief}_i},$$

[7] The ABC method has consistently been shown to generate accurate predictions in voluntary contribution games conducted *without* observability (Isler et al., 2021). Applying the ABC method to *No Observation*, the condition that lacks observability, confirms its accuracy in predicting contribution choices. On average, predictions are off by $-.21$ tokens from the true mean (on a possible range from -20 to 20), with a standard deviation of 4.79. Errors are assumed to be uncorrelated.

[8] The reason for this additional effect on top of effects transmitted through the ABC prediction has been the subject of speculation that remains inconclusive. We are agnostic about the underlying psychology. What matters for our purpose is that beliefs appear to exert independent influence - hence our inclusion as a separate explanatory variable.

the indirect effects via beliefs and the ABC prediction by

$$\text{Indirect Effect}_{\text{belief}\to\text{ABCpred}} = \frac{\partial\,\text{belief}_i}{\partial\,\text{treatment}_i} \times \frac{\partial\,\text{ABCpred}_i}{\partial\,\text{belief}_i} \times \frac{\partial\,\text{contribution}_i}{\partial\,\text{ABCpred}_i},$$

and the combined indirect effect is the sum of the two.

In our laboratory setting, contributions and beliefs are bounded and often concentrated (e.g., around focal values or due to distinct behavioural types). Moreover, if treatment affects social image concerns, contributions and beliefs may converge toward socially acceptable norms in some treatments, while in others, greater behavioural dispersion may emerge. To guard inference against departures from homoskedasticity and normality, we compute Huber-White robust standard errors.

Benchmark scenarios. The (direct) effects of being observed by a peer in treatments *Anonymous*, *Stranger* and *Friend* are captured by the coefficients $\delta 1$, $\delta 2$, and $\delta 3$, respectively. They represent the change in contribution behaviour among observed participants relative to the control condition, *No Observation*, net of potential belief-based effects caused by participants' anticipation of their group members' response to their observability. Depending on how social closeness influences peer effects among observed participants we expect the following relative magnitude of coefficients:

1. If peer effects are absent in our experiment, then $\delta_1 = \delta_2 = \delta_3 = 0$. This would imply that being observed has no *direct* effect on contribution choices.

2. If peer effects are present but orthogonal to social closeness to the observer, then $0 < \delta_1 = \delta_2 = \delta_3$. In this case, being observed has a direct positive effect on contribution behaviour, but the degree of social close-

ness plays no role in shaping decisions.

3. If the magnitude of peer effects *increases* with social closeness, then we should observe $0 < \delta_1 < \delta_2 < \delta_3$. This pattern would indicate that social closeness positively moderates the direct effect of being observed on contribution decisions. One interpretation is that observed participants are especially motivated to build or maintain a prosocial reputation vis-à-vis their friends.

4. If peer effects *decrease* with social closeness, then the expected pattern would be $\delta_1 > \delta_2 > \delta_3 > 0$. In this case, social closeness has a direct negative effect on contribution behaviour - the pattern consistent with the non-parametric analysis reported above. One interpretation is that the direct effect of closeness on contribution behaviour is determined by the stability of the individual's reputation. This would allow observed participants to feel more comfortable behaving selfishly in front of friends, knowing it is unlikely to damage their social image.

Estimated effects. Figure 1.4 plots the estimated coefficients δ_1, δ_2 and δ_3, capturing the direct effect of treatment on contribution behaviour, net of any indirect effects transmitted through beliefs. For the full regression results, including effects of beliefs on contributions, and of treatment on beliefs, see Appendix A.2, Table A.1. For an overview of treatment effects decomposed in direct, (combined) indirect and total effects refer to Appendix A.2, Table A.2.

Being observed by an anonymous person caused a *direct* increase in contributions by 2.69 ECU ($p < 0.001$), representing an economically large effect of a 31% increase relative to unobserved participants in *No Observation*. Being observed by a stranger produced an increase of 1.16 ECU, or 13% of the control

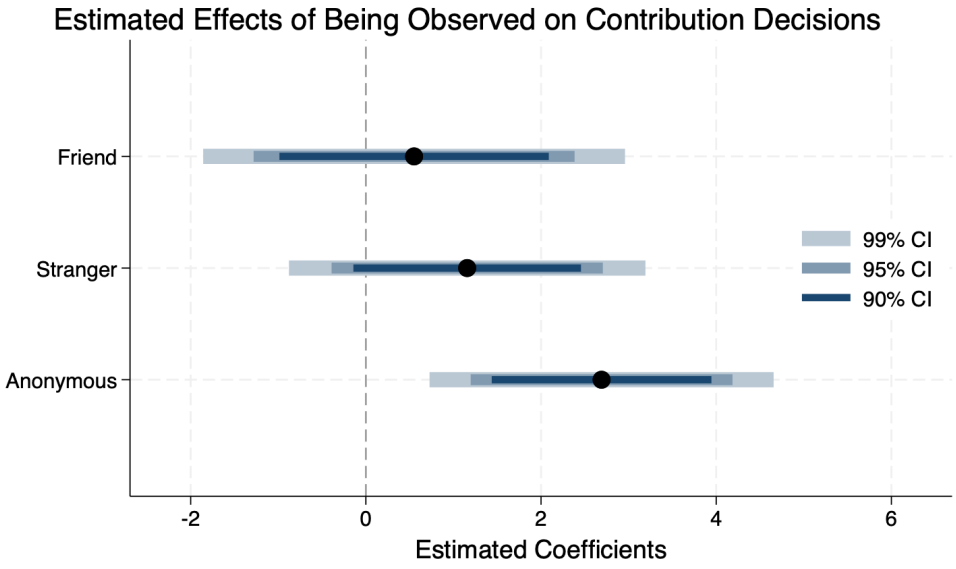


Figure 1.4: Plot shows estimated coefficients δ_1, δ_2 and δ_3 of Equation 1.1c in the structural equation model. Coefficients represent direct effects of being observed on contribution levels relative to contributions of unobserved participants in *No Observation*, measured in experimental currency units (ECU). The model controls for indirect treatment effects caused by treatment shifting beliefs about the average contributions of others, as well as for any effects of beliefs more generally and *ex ante* existing cooperativeness. Error bars indicate robust 90%, 95% and 99% confidence intervals.

mean, and in *Friend*, contributions increase by 0.55 ECU, or 6%[9]. In terms of direction and magnitude, the pattern thus points to the hypothesized benchmark scenario 4, with peer effects on observed participants decreasing with social closeness. Nonetheless, the coefficients for treatments *Stranger* and *Friend* are not statistically distinguishable from zero ($p = 0.143$ and $p = 0.556$, respectively). Post-estimation analyses reject the null of equal coefficients between *Anonymous* and both *Stranger* ($p = 0.045$) and *Friend* ($p = 0.021$), but not between *Stranger* and *Friend* ($p = 0.520$). Overall, these results support the interpretation that peer effects are strongest when the observer is anonymous and diminish as social closeness increases, but more precise estimation is needed to confirm the statistical significance of differences across different degrees of social closeness.

Regarding potential *indirect* effects of treatment on behaviour, we find that beliefs about the average contributions of others are an important determinant of contribution decisions, both directly and via the ABC prediction (see Table A.1 in Appendix A.2). However, this effect is *not* triggered by the observability of choices in any treatment: neither *Anonymous*, *Stranger*, nor *Friend* significantly affects beliefs, as indicated by the estimated treatment coefficients in the belief equation (left column of Table A.1). Accordingly, product-of-coefficients tests find no evidence of combined belief-based effects[10] (−0.07 ECU in *Anonymous* ($p = 0.938$); 0.82 ECU in *Stranger* ($p = 0.363$); 0.14 ECU in *Friend* ($p = 0.888$)), indicating that, within our experimental design, treatment-induced changes in beliefs do not substantially mediate contribution differences.[11] Thus, at least in this one-shot version of the game, expectations do not appear fully rational, in the sense that they do not fully incorporate treatment-induced be-

[9] Percentage increases reflect residual treatment effects, adjusted for beliefs and contribution types.

[10] We also find no evidence of significant effects transmitted through either of the two belief-based channels individually in any treatment.

[11] See Table A.2 in Appendix A.2 for a decomposition of treatment effects into direct, combined indirect, and total effects.

havioural responses of group members.

Model fit. The only valid fit indices under robust estimation are residual-based. The standardised root mean squared residual, which is based on the covariance residual, is $SRMR = 0.040$, indicating very good covariance fit according to even conservative conventional thresholds (Hu and Bentler, 1999). Equation level diagnostics reveal substantial heterogeneity in the explanatory power of the three endogenous equations. Equation 1.1c, which contains the direct effects, explains almost two thirds of the variance in contribution behaviour ($R^2 = 0.634$), Equation 1.1b explains about a third of the variance in ABC predictions ($R^2 = 0.363$), while almost none of the variance in beliefs is explained by Equation 1.1a ($R^2 = 0.005$). This is in line with our findings of the relevance of direct vs indirect effects in our experiment. Since the coefficient of determination (CD), indicating how much of the joint variance-covariance structure is explained, is multiplicative, it is driven down by the extremely poor explanatory power of Equation 1.1a, yielding an overall ($CD = 0.056$). It is important to note, however, that global fit is *not* our analytic objective. Our model is mechanism-targeted and motivated by theoretical considerations, aimed at identifying the direct effects of treatment on contribution choices net of any indirect channels running through beliefs. Since beliefs turn out empirically insensitive to treatment, the model necessarily fails to explain their variation. The resulting modest residual fit thus does not undermine the validity of the estimated direct treatment effects, but rather mirrors the empirical fact our model was designed to uncover.

We will next present the results of observing participants. Concluding remarks for both results sections will be provided in Section 1.5.

1.3.2 Observing others

There is a long-standing body of work across economics, evolutionary biology, psychology, and other disciplines documenting the widespread human tendency to conform to the behaviour of others. In different contexts, conformity has been interpreted as an equilibrium selection device in coordination problems (Gächter et al., 2023), adherence to norms (Bernheim, 1994; Young, 2015), adoption of successful strategies in environments of incomplete information (Bikhchandani et al., 1992, 1998), or a response to social sanctions or rewards (Dvorak et al., 2025). But why, if at all, should we expect participants in the role of observers to conform to the contributions they observe when making their own decisions? And how does social closeness to the observed participant interfere with such conformity? Participants aiming solely to maximise their monetary payoff from the experiment should keep their entire endowment for themselves, regardless of the choice they observe or how close they are to the observed person. Since instructions carefully explained the structure of the game and all participants successfully answered a series of control questions, we can reasonably assume that the payoff-maximising strategy was common knowledge. Therefore, acquiring information about the materially optimal choice plausibly did not play a significant role in shaping observers' decisions. Moreover, since observers themselves were unobserved, reward or punishment mechanisms were absent by design. Ruling these mechanisms out, our experiment is thus well suited to study whether observed contribution choices influence observers' own decisions in ways that cannot be attributed to monetary incentives alone.

Hypotheses

Indirect paths. A common motivation in voluntary contribution games is conditional cooperation - contributing more when expecting others to do the same

(Fischbacher et al., 2001; Thöni and Volk, 2018; Chaudhuri et al., 2006). In the context of our experiment, observers may update their expectation about their group members' contributions in response to the decisions made by their peer, which in turn may shape their own decisions. In contrast to observed participants, where belief-based mechanisms arose due to *others* interacting with their peers, observing participants may update their beliefs based on observing a peer themselves. Put differently, had the experimental design ruled out peer influence on group members, belief-based effects would vanish among observed, but remain unaffected among observing. Therefore, without taking a hard stance on the labelling of belief-based effects, we consider effects caused by others interacting with peers - the case of observed participants - as a confounding mechanism, while effects caused by a participant's own exposure to peer influence - the case of observers - as a mechanism that plausibly falls within a useful definition of peer effects. Empirically, belief-based effects among observers should manifest in beliefs about group members' choices influencing contributions, and observed contributions influencing beliefs.

Direct path. Besides monetary incentives, participants may also care about *im*material qualities of their decision. While our experiment is not designed to study the psychological mechanisms underlying such preferences, these likely involve how one's own contribution compares to those of others, evoking social identity, or whether one's choice falls within what is considered acceptable behaviour.

With regard to motives related to social identity, a large body of literature demonstrates that individuals tend to align choices with in-group members.[12] Thus, if observers' contribution choices are shaped by social identity motives, we ex-

[12]For example, Benjamin et al. (2010) document the effect of ethnicity-primes on time and risk preferences, and Benjamin et al. (2016) find a causal link of religious identity on behaviour in voluntary contribution games. For excellent overview articles on how identity shapes (economic) behaviour, we refer to Charness and Chen (2020) and Shayo (2020).

pect them to be more inclined to conform to the choices of their friends than to strangers or anonymous participants.

Potential mechanisms related to preferences for norm conformity and why they may cause observing participants to make contributions more similar to their friends than to strangers or anonymous participants are more involved. In the following, we shall provide an illustrative example.

Consider a scenario in which each player receives a private signal about the normatively right contribution. To the extent that observed players care about norm conformity, their choices will carry information about their perception of the norm. This allows observing players to extract information from the observed choice and combine it with their own private signal to form a weighted posterior perception of the norm. Importantly, the weight placed on this extracted information will depend on its accuracy. Thus, assuming that accuracy increases with social closeness - e.g., because players are better able to interpret the motives behind a friend's contribution than that of a stranger - players will put more weight on the observed choice when it comes from a friend, aligning their contributions.

Results

Analogously to Section 1.3.1, we begin with a short description of the data, shown in Figure 1.5.

Participants predominantly fall into the category of conditional cooperators. Despite random treatment allocation and a χ^2 test not rejecting the null of equal type distribution among participants in *No Observation* and observers in *Anonymous*, *Stranger* and *Friend* ($p = 0.12$), there are numerically fewer free riders in *Anonymous* and *Stranger*. To account for this, we control for contribution types

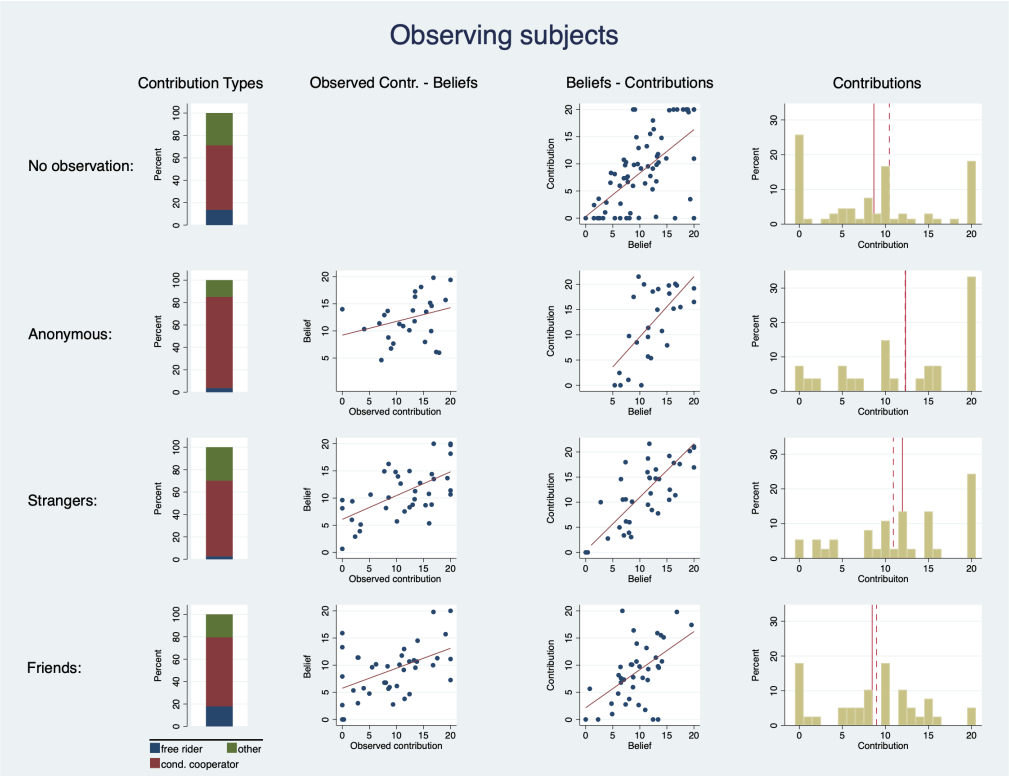


Figure 1.5: Overview - observing subjects. Left: distribution of contribution types based on strategy profile from Phase 1. Middle-left: Scatterplot of beliefs about group members' contributions and observed contributions, with fitted regression line. Middle-right: Scatterplot of individual contributions and beliefs about group members' contributions in Phase 2, with fitted regression line. Right: Histogram of contributions in Phase 2, with mean contributions (solid line) and mean beliefs about group members' average contribution (dashed line).

in all subsequent analyses.

We observe a positive association between observed contributions and participants' belief about average contributions in their own group. Moreover, participants' beliefs positively correlate with their contributions. Together, this pattern is consistent with observed contributions affecting observers' choices not only *directly*, but also *indirectly* by shaping expectations of the contributions of their group members.

The histograms of contributions show similar distributions in *Friend* as in *No Observation*, including the prevalence of free riding, while contributions in *Anonymous* and *Stranger* are skewed to the left. Note, however, that we are not interested in their contributions *per se*. We are interested in how observed choices influence observing participants' own behaviour, and whether this influence varies across treatments.

Estimation strategy. To explore this influence and separate its underlying mechanisms in direct and indirect (belief-based) channels, we simultaneously estimate the following system of structural equations:

$$\begin{aligned}\text{Belief}_i = \theta_0 &+ \theta_1 \text{Contribution}_i^{obs} + \theta_2 S_i + \theta_3 F_i + \theta_4 (\text{Contribution}_i^{obs} \times S_i) \\ &+ \theta_5 (\text{Contribution}_i^{obs} \times F_i) + \varepsilon_i^{(d)} \end{aligned} \tag{1.2a}$$

$$\text{ABCpred}_i = \kappa_0 + \kappa_1 \text{Belief}_i + \varepsilon_i^{(e)} \tag{1.2b}$$

$$\begin{aligned}\text{Contribution}_i = \tau_0 &+ \tau_1 \text{Contribution}_i^{obs} + \tau_2 S_i + \tau_3 F_i + \tau_4 (\text{Contribution}_i^{obs} \times S_i) \\ &+ \tau_5 (\text{Contribution}_i^{obs} \times F_i) + \tau_6 \text{Belief}_i + \tau_7 \text{ABCpred}_i + \varepsilon_i^{(f)} \end{aligned} \tag{1.2c}$$

Our model treats treatment *Anonymous* as the benchmark condition. Treatments

Stranger and *Friend* are denoted by the dummy variables S_i and F_i. The contribution choice observed by an individual - that is, the choice made by their peer - is given by $Contribution_i^{obs}$, their expectations about their group members' average contribution are denoted by $Belief_i$, and their ABC prediction, as discussed in Section 1.3.1, by $ABCpred_i$.

The model is specified identically to Equation 1.1, the structural model of observed participants, but extends it by allowing for the observed contribution to affect beliefs and contribution choices. If these effects vary across treatments either in the belief- or the contribution equation, this will be captured by the interaction terms. We compute Huber-White robust standard errors to account for potential heteroskedasticity or deviations from normality.

Our aim is to separate direct from indirect belief-based effects that observed contributions may exert on participants' own choices. The direct effect runs from observed to own contribution, holding beliefs (and hence also the ABC prediction) constant. Formally, it is given by

$$\left.\frac{\partial \,\text{contribution}_i}{\partial \,\text{contribution}_i^{obs}}\right|_{\text{belief}_i,\ \text{ABCpred}_i} = \tau_1 + \tau_4 S_i + \tau_5 F_i.$$

The combined indirect (belief-based) effect, caused by the observed contribution influencing beliefs and hence the observers' own contributions directly, and indirectly via the ABC prediction, is given by

$$\frac{\partial \,\text{belief}_i}{\partial \,\text{contribution}_i^{obs}} \times \left[\frac{\partial \,\text{contribution}_i}{\partial \,\text{belief}_i} + \frac{\partial \,\text{ABCpred}_i}{\partial \,\text{belief}_i} \times \frac{\partial \,\text{contribution}_i}{\partial \,\text{ABCpred}_i}\right]$$

$$= (\theta_1 + \theta_4 S_i + \theta_5 F_i)(\tau_6 + \kappa_1 \cdot \tau_7)$$

Note the relevance of including the ABC prediction Equation 1.2c: Especially

among friends, pre-existing similarities in cooperativeness may yield similarities in contribution choices even in the absence of observability of choices. Since the ABC prediction is based on *ex ante* elicited contribution preferences, its inclusion as a covariate controls for such spurious similarities in contributions and allows us to attribute residual similarities to the fact that one friend observes the other's choices.[13]

Benchmark scenarios. Depending on how social closeness *directly* shapes peer effects on observing participants, one of the following patterns should emerge.

1. If observed contributions do not influence observers' decisions net of belief-based effects, then: $\tau_1 = \tau_1 + \tau_4 = \tau_1 + \tau_5 = 0$.

2. If direct (positive) influence exists, but without systematic association with social closeness, then: $0 < \tau_1 = \tau_1 + \tau_4 = \tau_1 + \tau_5$. This scenario would suggest a tendency for observers to mimic the contribution choices of their peers, for example, because they use it as an indication of appropriate behaviour, no matter how close they are.

3. If direct effects increase with social closeness, then we expect: $0 < \tau_1 < \tau_1 + \tau_4 < \tau_1 + \tau_5$. A possible interpretation of this pattern would be that observing participants regard socially close peers as a more reliable reference for appropriate behaviour than socially distant ones. Alternatively, social identity motives evoking imitation may be triggered by friends' behaviour but not by strangers' or anonymous participants'.

[13] We examine whether friends have more similar contribution preferences than non-friends by comparing the contribution profiles as elicited in Phase 1, the strategy method public good game which contains no element of observability (see Appendix A.1 Figure A.2). We find that friends have (slightly) more similar contribution preferences than non-friends, which underscores the necessity of such controls in our model.

4. In principle, direct peer effects may also decrease with social closeness. In this case, the following pattern should arise: $\tau_1 > \tau_1 + \tau_4 > \tau_1 + \tau_5 > 0$. While the interpretation of this pattern seems theoretically unclear, it shall be mentioned here for completeness.

Beliefs appear twice in Equation 1.2c - explicitly and implicitly in the ABC prediction. Consequently, our model allows for two indirect channels from treatment to observers' contribution choices. Since this specification is motivated by findings on belief-based effects in previous literature which are not yet fully understood and are of limited relevance to our research question, we focus on the combined effect of the two belief-based channels. Depending on how they are affected by social closeness, one of the following patterns should arise.

1. If there are no indirect effects, then: $\theta_1(\tau_6 + \kappa_1 \cdot \tau_7) = (\theta_1 + \theta_4)(\tau_6 + \kappa_1 \cdot \tau_7) = (\theta_1 + \theta_5)(\tau_6 + \kappa_1 \cdot \tau_7) = 0$.

2. If indirect effects do shape contribution choices but are independent of social closeness, then: $0 < \theta_1(\tau_6 + \kappa_1 \cdot \tau_7) = (\theta_1 + \theta_4)(\tau_6 + \kappa_1 \cdot \tau_7) = (\theta_1 + \theta_5)(\tau_6 + \kappa_1 \cdot \tau_7)$. It would imply that the observation of a peer's choice does influence contribution choices by shaping beliefs, but not more (or less) so depending on social closeness.

3. If indirect effects become more pronounced with social closeness, then we should see: $0 < \theta_1(\tau_6 + \kappa_1 \cdot \tau_7) < (\theta_1 + \theta_4)(\tau_6 + \kappa_1 \cdot \tau_7) < (\theta_1 + \theta_5)(\tau_6 + \kappa_1 \cdot \tau_7)$. In this case, the effect on observers' beliefs is stronger when the observed contribution was made by a person who is socially closer than observed individuals in treatment *Anonymous* (captured by θ_4 for *Stranger*, and θ_5 for *Friend*).

4. Finally, if indirect peer effects decrease with social closeness, we should observe: $\theta_1(\tau_6 + \kappa_1 \cdot \tau_7) > (\theta_1 + \theta_4)(\tau_6 + \kappa_1 \cdot \tau_7) > (\theta_1 + \theta_5)(\tau_6 + \kappa_1 \cdot \tau_7) > 0$.

Estimated effects. Figure 1.6 shows the estimated responsiveness of observing participants to the contribution decisions of their peers decomposed into direct, indirect and total effects. For the full regression table please refer to Appendix A.2 Table A.3. An overview of treatment effects decomposed in direct, (combined) indirect and total effects is provided in Appendix A.2, Table A.4.

While none of the direct effects we estimate for the three treatments is individually distinguishable from zero, we find that responsiveness to peer contributions is significantly greater when the peer is a friend compared to when the peer is anonymous. The estimated difference in slopes between treatments *Anonymous* vs. *Friend* is 0.41 ECU ($p = 0.03$), implying that if, for instance, the observed contribution varies by 10 ECU, the average responsiveness differs by 4.1 ECU. In sum, friendship alters the direct behavioural contagion observed under anonymous interactions.

In terms of belief-based channels, we find clear evidence for observed contributions to indirectly influence observing participants' choices by changing their expectations about the contributions of their group members. For all treatments, Wald tests on the product of coefficients reject the null that the point estimates are equal to zero (*Anonymous*: 0.24, $p = 0.061$; *Stranger*: 0.41, $p < 0.001$; *Friend*: 0.35, $p = 0.006$). Consistent with the hypothesized Scenario 2 outlined above, we do not find differences across treatments, suggesting that social closeness does not moderate belief-based effects on observers' contribution behaviour.

The total effect of both direct and belief-based influences of observed peers on observers' behaviour reflects the individual effects of each channel. Specifically, the estimated total response to a unit increase in observed contributions equals 0.04 ($p = 0.877$) in *Anonymous*, 0.43 ($p = 0.002$) in *Stranger*, and 0.56 ($p < 0.001$) in *Friend*. These results constitute economically meaning-

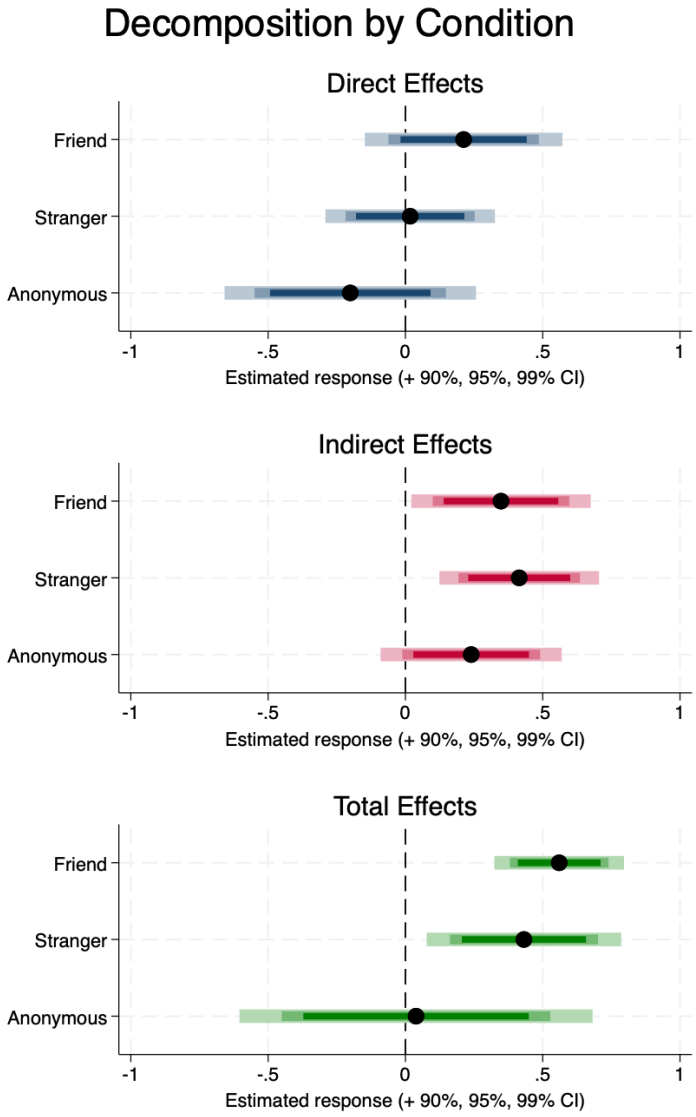


Figure 1.6: Plots show estimated direct, indirect (belief-based) and total responsiveness of observing participants to a unit increase in the observed contribution made by a peer, derived from a structural equation model estimated with robust standard errors. Direct effects capture the influence of observed contributions holding beliefs about the contributions of group members constant (computed by the sum of the structural coefficient on the observed contribution and its interaction with the treatment dummies). Indirect effects correspond to changes in contributions caused by updated expectations about the average contributions of group members (given by the product of coefficients and their interactions with treatment dummies) representing the sum of the paths *Treatment* → *Belief* → *Contribution* and *Treatment* → *Belief* → *ABCpred* → *Contribution*. Total effects are given by the sum of direct and indirect effects. The model controls for *ex ante* existing differences in beliefs and cooperativeness. Point estimates are displayed with 90%, 95% and 99% confidence intervals and reported separately for each treatment.

ful evidence that peer influence is critically shaped by social closeness, with differences becoming especially pronounced when comparing anonymous participants and friends ($p = 0.047$).

Model fit. Goodness-of-fit analyses yield a standardised root mean squared residual of $SRMR = 0.061$, indicating low average discrepancy between predicted and observed covariances, and thus good overall fit. Moreover, our model accounts for approximately 41% of the variances and covariances ($CD = 0.408$). The explanatory power of the three endogenous equations is considerable: Equation 1.2c explains nearly two thirds of the variance in contribution behaviour ($R^2 = 0.645$), Equation 1.2b explains more than half of the variance in ABC predictions ($R^2 = 0.527$), and Equation 1.2a explains one third of the variance in beliefs ($R^2 = 0.358$). The stark contrast in explanatory power between Equation 1.2a (modelling the beliefs of observers) and Equation 1.1a (modelling the beliefs of observed participants) mirrors the finding that observers' beliefs are substantially shaped by their observation of a peer's choice, which, by design, is absent in the context of observed participants.

Discussion

Observability causes similarity of contributions in *Friend*. The (total) responsiveness to observed contributions as estimated by Equation 1.2 (+0.04 in *Anonymous*, +0.43 in *Stranger*, and +0.56 in *Friend* per unit increase of the observed contribution) implies that choices become more similar with social closeness under observability. By including the ABC prediction as a covariate in Equation 1.2c, we control for *ex ante* existing similarities in contribution preferences. This enables us to causally attribute the (residual) similarity in choices

to observability, as opposed to, e.g., homophily.[14]

We further support our claim that the estimated similarity of friends' choices is caused by the experimentally induced observability and not by similarity in cooperativeness by contrasting how similar choices are when there is observability with when there is not. To do so, we use the unconditional contribution choices of the strategy method in Phase 1, which lacked observability, as a counterfactual to the contributions made in Phase 2. Both phases were identical in all other experimental parameters, and featured the same pairs of friends. Figure A.3 in Appendix A.1 plots the similarities of friends' choices with (Phase 2) and without observability (Phase 1). Choices are noticeably more similar with observability. A Wilcoxon signed-rank test rejects the null of equal distributions ($p = 0.001$) across the two phases. A placebo test on the similarities of friends' choices across Phase 1 and Phase 2 in *No Observation*, that is, when *both* phases lack observability, confirms its causal role: a Wilcoxon signed-rank test now detects no difference in similarities ($p = 0.49$).

Direction of influence. Thus far, we have implicitly treated the results as evidence that observing participants "follow" the choices they observe more closely when they were made by a friend instead of an anonymous participant. We add the caveat that this interpretation deserves a closer look. Our assumption originates from the fact that observers do see their observed peers' choices, but not vice versa. However, the direction of influence could - in principle - also go the other way. This would require observing participants to be unaffected by what they see and observed participants to correctly anticipate their observer's contribution and conform to it.

The identification challenge originates from the fact that, while we exogenously

[14]See Appendix A.1 Figure A.2 for a comparison of *ex ante* contribution preferences of friends vs. non-friends.

control social closeness between peers through treatment assignment, the contributions observers see are endogenously determined by observed participants. To substantiate our claim that observed contributions influence observers, and not the other way around, we thus adopt the following two distinct strategies.

1. The estimates from our structural equation model indicate that higher social closeness (measured by our treatment manipulation) is associated with higher similarity of observed and observing participants' choices. A Jonckheere–Terpstra test confirms this result. It detects a highly significant trend of increasing similarities of contributions from treatment *Anonymous* to *Stranger* to *Friend* ($p < 0.001$). To examine if the association of social closeness and similarity of choices is indeed driven by observing, or rather by observed participants, we draw on the IOS scores as an alternative measure of social closeness. Since IOS scores measure *subjective* closeness to another person, there is variation in this measure within observation pairs. Thus, we can check if this association replicates (better) on the basis of the ratings of observed or on those of observing participants. We find the same highly significant trend with regard to the IOS ratings of observing participants ($p < 0.001$) as we did with regard to treatments, but *not* with regard to the ratings of observed participants ($p = 0.171$). In other words, the choices of participants who form an observation pair become significantly more similar when the observing participant feels close, but not when the observed participant does. Thus, there is no indication in the data that the similarity of choices is caused by observed participants conforming to the anticipated choices of their observers. This is particularly remarkable, considering that high IOS ratings are associated with how well people know each other, and thus should correlate with observed participants' ability to anticipate observers' choices.

2. We have previously shown that observability causes similarity of choices in treatment *Friend*. In other words, friends' choices are similar in Phase 2, but

not in Phase 1. Therefore, at least one of them (observed or observing) must have changed their contribution in Phase 2 relative to Phase 1, and thus be responsible for the similarity in Phase 2. Based on this reasoning, we examine how observed and observing participants change their contribution choices in Phase 2 relative to Phase 1 against the benchmark of how participants in treatment *No Observation* change their contributions from Phase 1 to Phase 2 *in the absence* of observability.[15] A two-sample Wilcoxon rank-sum test does not find observed participants in *Friend* to change their contributions between Phase 1 and Phase 2 more than participants in *No Observation* ($p = 0.355$, one-sided). In contrast, observing participants in *Friend* make significantly larger changes than participants in *No Observation* ($p = 0.048$, one-sided).

Based on these results, we conclude that the observed pattern - where contribution choices become more similar with social closeness under observability, especially in treatment *Friend* - is likely driven by the behaviour of observers.

1.4 Discussion

To study social closeness as broadly and realistically as possible, our experiment involved studying peer effects between friends. Accordingly, the recruitment procedure required participants to bring a friend who had to be willing to commit several days prior to the session. This policy caused unexpectedly low sign-up rates and, because cancellation of one participant automatically implied that their friend could not be allowed into the experiment either, a substantial proportion of no-shows. This significantly increased the duration and costs of our study and resulted in a lower-than-expected number of observations. As a

[15] We use *No Observation* as a benchmark because virtually all studies on public good games find that contributions decline over time. *No Observation* serves as an ideal counterfactual scenario featuring all potential motives for changes in contribution choices between Phase 1 and Phase 2 - except for those triggered by observability.

consequence, the estimates from the structural equation models partly exhibit wide confidence intervals. Therefore, despite clear patterns in the point estimates, they do not always achieve conventional levels of statistical significance. Based on this experience, we recommend that researchers planning studies involving similar recruitment procedures account for low sign-up rates and elevated no-show rates, for instance, by allowing for significant overbooking and waitlists.

1.5 Conclusion

To the best of our knowledge, our experiment is the first to provide evidence that social closeness shapes peer effects. We consider both types of peer effects: behavioural changes caused by being observed, and by observing a peer. Importantly, the combination of a novel experimental design and structural equation modelling allows us to separate underlying mechanisms into direct effects on behaviour and belief-based indirect effects, which are of particular relevance for cooperation problems.

With regard to effects triggered by being observed by a peer, we replicate the general finding of the literature that people become more prosocial when feeling observed. However, we show that this effect is crucially shaped by their social closeness to the observer. Specifically, it holds vis-à-vis anonymous peers (the typical condition in economic experiments) and vanishes with increasing levels of social closeness. For example, the behaviour of participants who are observed by a friend is indistinguishable from unobserved participants in the control condition. This is remarkable, because a positive relationship between image concerns and social closeness seemed intuitive. Instead, the observed behavioural pattern is consistent with participants trading off social image con-

cerns against selfishness, where the opportunity cost of selfishness is decreasing with reputational ambiguity and hence social closeness. In line with previous literature, expected behaviour of others has a large effect on contribution choices. However, we do not find evidence for belief-based treatment effects. Thus, participants in our experiment fail to account for the treatment-induced changes in behaviour of their group members.

With regard to effects caused by observing a peer, we find significant differential influence of peer behaviour on contribution choices with respect to social closeness. Specifically, the estimated slope capturing the relationship of observed and own contributions is significantly higher when the observed peer is a friend instead of an anonymous peer. Moreover, we find significant belief-driven peer influence on observing participants. However, this effect is not associated with social closeness. The total effect of direct and indirect channels is significantly shaped by social closeness. In sum, our findings indicate that when information about the behaviour of others is available, social closeness determines whether people conform or not. The closer the decision maker feels to the observed individual, the stronger the conformity effect.

Chapter 2

Social Influence on Moral Judgements

Abstract

How moral judgements are shaped by social influence, despite being based on fundamental moral values, remains a long-standing question in the social sciences. This study presents data from an interactive online experiment documenting substantial social influence via two distinct channels: First, providing information about the moral opinions of others induces significant intrinsic conformity - a shift of moral judgements towards the majority opinion. Second, exposure to social evaluation causes strategic distortions of publicly visible moral judgements, resulting in even higher conformity.

This chapter is based on joint work with Fabian Dvorak, Urs Fischbacher, and Katrin Schmelz. We thank Robin Cubitt, Simon Gächter, Felix Holzmeister, Botond Kőszegi, Yana Litovsky, Seung-Keun Martinez, Devesh Rustagi, Silvia Sonderegger, and Florian Zimmermann, as well as seminar participants in Bonn, Innsbruck and Nottingham, and participants at the CCC conference in Amsterdam, the Spring Summit in Innsbruck, the IAREP-SABE Meeting in Nice, and the ESA World Meeting in Lyon, for helpful comments. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2117 - 422037984. We also gratefully acknowledge financial support from the CeDEx Research Centre at the University of Nottingham School of Economics.

2.1 Introduction

Moral dilemmas are ubiquitous in our daily lives, with far-reaching consequences for society. Market interactions (Falk and Szech, 2013; Bursztyn et al., 2019b; Pruckner and Sausgruber, 2013; Kirchler et al., 2016; Falk et al., 2020), political outcomes (Haidt, 2007; Graham et al., 2009; Haidt, 2012; Enke, 2020a), and cooperation in social dilemmas (Gintis et al., 2008; Henrich et al., 2010; Tomasello, 2016; Curry et al., 2019) have been shown to be significantly shaped by morality. Moral dilemmas also lie at the heart of recent political controversies: Is it right to deny a woman's right to bodily autonomy to protect unborn life? Is it right to engage in illegal activities to call attention to the devastating consequences of climate change? To send lethal military equipment to war zones so that countries under attack can defend themselves against invasion? Or to compromise people's physical integrity to save human lives by making vaccinations mandatory?

All of these problems concern different moral norms and are situated in vastly different contexts. But they share the same logic: obeying one moral rule implies violating another. Hence, they constitute sacrificial moral dilemmas (Foot, 1967; Thomson, 1976). This study examines whether people's judgements on such moral problems are susceptible to social influence.

While there is evidence that decisions in moral dilemmas can be shaped by the provision of information on the moral beliefs of others (Kundu and Cummins, 2013; Rom and Conway, 2018; Bostyn and Roets, 2017), morality is also commonly understood as anchored in fundamental moral values (Feather, 1995; Hitlin and Piliavin, 2004; Rohan, 2000; Schwartz, 2012; Tetlock, 1986), supporting the view that moral judgements are stable. Here we provide clean evidence that moral judgements are significantly shaped by the opinions of others, and uncover the mechanisms behind this dynamic.

Methodologically, pinpointing the causal effect of social influence on people's moral judgements is challenging. Clean identification requires a within-subject comparison of judgements on one and the same moral problem, once made in private and once under social influence. This involves two main challenges. First, simply confronting people with the same moral question twice, first in social isolation and then again under (whatever kind of) social influence may bias estimates downward due to concerns for consistency (Falk and Zimmermann, 2017; Fehrler and Hughes, 2018; Soll and Larrick, 2009). Second, while between-subject comparisons may at first glance seem like a natural remedy to this problem, they are, however, inadequate for identifying the effects of social influence. To see why, note that social influence may induce changes in moral views in two opposing directions: towards and away from the opinion of others. Thus, if both happen at the same time, which *a priori* cannot be excluded, the distribution of moral judgements may look spuriously similar before and after, or with and without social influence, as individual changes may cancel out on the population level.

We use an interactive online experiment which advances the design by Dvorak et al. (2025) and exploits the decision-theoretic principle that preferences should satisfy transitivity. This allows us to obtain both judgements - private and socially influenced - on the same moral problem and from the same participant, while circumventing potential concerns for moral consistency. It further permits distinguishing between *genuine* changes in moral judgements and *strategic* (mis)reporting, a concern commonly associated with social desirability bias.

The results document that social influence significantly affects people's judgements through two distinct mechanisms: intrinsic conformity, which reflects changes in moral preferences toward the majority judgement, and strategic conformity, which aims to attract praise or avoid punishment from others.

The observed intrinsic conformity is remarkable because there are no information asymmetries between participants and hence there is no (obvious) rationale for social learning. It supports the idea that individuals adopt the moral values of others when they are common in a population (see Boyd and Richerson, 1988; Bowles and Choi, 2013; Denton et al., 2020), and is robust across all of the twelve moral problems we consider - from extramarital affairs to triage in emergency situations.

Strategic conformity arises when participants strategically distort their publicly expressed moral values to attract reward or avoid punishment from other participants. Our findings suggest that it substantially reinforces people's tendency to intrinsically conform to the majority view, and connects directly to the previous literature on strategic conformity in *a*moral domains (Festinger, 1953; Kelman, 1961; Allen, 1965; Amabile et al., 1990; Bernheim, 1994; Shalley and Perry-Smith, 2001).

More generally, both observed patterns of conformity connect to a long-standing literature (Boyd and Richerson, 1988; Bowles and Choi, 2013; Denton et al., 2020; Apesteguia et al., 2007; Schlag, 1998, 1999; Vega-Redondo, 1997; Bernheim, 1994), initiated by seminal contributions of social psychologists (Sherif, 1935; Asch, 1951), which highlights the role of social influence for human decision making (see, for example, Asch (1955) or Cialdini and Goldstein (2004)).

We believe that our findings are relevant for a wide range of societal domains. They suggest that when moral issues enter the public sphere, others' opinions may significantly shape how people view them, or, when expressing their own opinions, they may misreport them to match the dominant view. For example, political debates often centre on moral dilemmas, and people form judgements under the impression of public discussions and of opinions encountered in traditional and social media.

Moreover, the conformity patterns observed in this study directly inform current discussions - both within and beyond academia - on the role of social media in shaping public discourse and society as a whole. With the rise of social media, the potential of social influence on people's opinions has reached unprecedented levels. Our results suggest that the opinion dynamics on social media is only inadequately captured by the fact that people's moral preferences are malleable and sensitive to information about the moral judgements of others. By their very definition, social media platforms not only make the opinions of others omnipresent for passive exposure, but also encourage liking, sharing, or other forms of active expression, and thus open doors for intrinsic adoption of, but also strategic adaptation to, circulating views. Since both effects exhibit significant patterns of conformity, they jointly carry the potential of amplifying the convergence of opinions in online communities, thereby promoting polarisation when different communities are exposed to different content.

With regard to belief formation based on selected information more generally, our results complement previous findings that people take what they see as an unbiased signal about the true state of the world (Enke, 2020b). While these are mainly applicable to *objective* facts towards which rational beliefs should, but factually do not (fully) converge, our study shows that similar patterns also apply to moral judgements, and hence to the domain of *subjectivity*.

Lastly, our results also inform recent strategies to deal with unavoidable ethical questions surrounding modern technology, such as autonomous cars. When harmful or even deadly accidents cannot be avoided, and saving one person may imply killing another, computers are often programmed to decide based on the aggregated judgements of lay people (Awad et al., 2018). If public debates or media coverage carry the potential to significantly shift people's moral sentiments, this may invalidate such aggregation processes and therefore the representativeness of previously collected patterns and the moral soundness of AI

technology.

2.2 Design and Procedure

2.2.1 Overview of the design

To isolate social influence from other potential drivers of changes in people's moral judgements, we advance the design by Dvorak et al. (2025). It further enables us to obtain two judgements on the same moral problem made by the same participant, one made in private and one while being exposed to social influence. To avoid potential concerns for moral consistency, it exploits the decision-theoretic principle that preferences should satisfy transitivity.

Specifically, we elicit judgements between options A versus B and B versus C. If a participant prefers A over B and B over C, we can, to the degree that preferences satisfy transitivity, infer that the participant prefers A over C - without having directly asked for a judgement between A and C. We also measure the strength of preference, which allows us to predict a preference in cases where, for example, A is preferred over both B and C, but only moderately over B and strongly over C. In this case, we would predict that B is preferred over C.

To empirically assess the intransitivity assumption, we implement the control condition *No info*. It consecutively elicits moral judgements on all three possible pairwise comparisons and thus allows us to obtain a benchmark level of discrepancy between the judgement on A versus C as inferred by transitivity and the directly measured one.

In all other treatments, we introduce different forms of social influence.

In treatment *Only info*, participants learn about two other participants' moral

views on problem A versus C prior to making a judgement themselves. Since the first two judgements are made under identical conditions as in *No info*, any increase in the probability of intransitivity relative to *No info* can be causally attributed to the provision of information about the moral views of others.

Finally, treatments *Reward* and *Punishment* create strategic incentives for participants to distort their stated moral judgement. As in *Only info*, participants learn about the views of two other participants on the moral problem A versus C. In addition, they are informed that the three stated views - theirs and those of the two other participants - will be disclosed to another participant who is tasked with rewarding (treatment *Reward*) or punishing (treatment *Punishment*) one of the three participants for their moral view. Comparing the probability of intransitivity in *Reward* or *Punishment* with the one we see in *Only info* reveals how participants distort stated moral judgements when they are made public and get evaluated.

2.2.2 Experimental conditions

In total, we implemented four experimental conditions: *No info*, *Only info*, *Reward*, and *Punishment*. We will first describe the details of *No info*, which serves as our control condition and constitutes the baseline of all other treatments. Subsequently, we will introduce treatments *Only info*, as well as *Reward* and *Punishment*.

No info

We confront all participants with a series of twelve moral trilemmas. These are descriptions of morally ambiguous scenarios similar to those in the opening paragraph and discussed by Thomson (1976) or the famous trolley car dilemmas

introduced by Foot (1967). Two notable differences are that our trilemmas offer three options instead of two, and that great emphasis has been placed on making them depict realistic situations. The topics range from extramarital affairs and forced blood transfusions to selection criteria for the recipients of life-saving organs. For illustration, consider the following trilemma from our experiment: The entire set of moral problems which we used can be found in Appendix B.2.

> Emil is a doctor and works in an organ transplant clinic in which there is an urgent need for donor lungs. Since there is only one donor lung available, Emil needs to decide who is getting the organ and who is inevitably going to die.
>
> According to your personal moral conviction: Who shall the donor lung be given to?
>
> A. A 20-year-old smoker, who is not an organ donor himself.
>
> B. A 45-year-old non-smoker, who is not an organ donor himself.
>
> C. A 70-year-old non-smoker, who is an organ donor himself.

The moral ambiguity is induced by a conflict of moral principles.

Importantly, out of the three options participants are only shown two at a time, for example A vs. B (see decision screen B.1 in Appendix B.5). Their task is to indicate which of the two they find morally superior, as well as the strength of their preference on a continuous scale. The experiment ends when all participants have been confronted with all three possible pairwise comparisons from each of the twelve moral trilemmas.[1]

The key thing to note is that after the first two sets of pairs, when, for example, a participant has stated their preference between option A vs. B and option B vs.

[1] In addition to the twelve trilemmas, we confronted participants with three versions of the classic sacrificial (trolley) dilemmas (see Appendix B.3) after the twelve trilemmas. The idea was to have a potential reference, if needed. We do not use them in our main analysis.

C, we can infer this participant's preference on the remaining pair A vs. C by assuming transitivity. We also elicit the strength of preference on a continuous scale. Thus, if the first two directly elicited moral judgements happened to be A > B and A > C, we can use the respective strengths of these preferences to infer the preference on B vs. C. This gives us two measures of the same moral preference A vs. C: one as inferred from the previously made choices between A vs. B and B vs. C, and one directly elicited choice between A vs. C.

Note that comparing the inferred with the explicitly stated preference on A vs. C constitutes a direct test of the transitivity assumption. Note further that, as we will explain next, in treatments *Only info*, *Reward*, and *Punishment*, social influence is implemented between the first two and the remaining third choice. Thus, in these treatments, we infer the preference on A vs. C from two choices that were made *before*, and then we directly elicit the preference on A vs. C *after* social influence was exerted. Hence, and this is the key idea of our design, while in *No info* the inferred and the explicitly stated preference can differ only when preferences violate transitivity, in treatments *Only info*, *Reward*, and *Punishment* they may also differ if the moral preference has changed due to social influence. Thus, *No info* enables us to identify the effect of social influence on moral preferences, as it quantifies the extent to which intransitivity causes mismatches of inferred and explicitly stated preference and hence allows us to causally ascribe any residual mismatches in *Only info*, *Reward*, and *Punishment* solely to social influence.

Only info

In treatment *Only info* we inform participants about the moral judgements of two other participants to examine whether their moral views change upon learning about those of others. The implementation works as follows.

We randomise participants into groups of three. As in *No info*, we then (individually) ask them to make judgements between the pairs of options (see Figure B.1 in Appendix B.5). This process continues until each participant has stated a preference on two out of the three pairs from each moral problem. Importantly, pairs of options are allocated to participants in a way such that the remaining pair is a different one for each group member. For example, if a given participant has not yet stated a preference on A vs. B, their two group members already have. Even though it had no direct relevance in *No info*, for reasons of consistency this specific order in which the pairs of options come up was held constant across all four experimental conditions.

As in *No info*, the experiment proceeds by confronting participants with the remaining pairs of options. Again, we ask them to state their moral preference. This time, however, along with showing the two options, we also inform them about the moral judgements of their two group members as stated in the previous phase (see Figure B.2 in Appendix B.5).

We call the judgements obtained in this way 'socially informed'. Accordingly, we call the moral preferences which we infer based on the respective two previous judgements *socially uninformed*. Again, note that these are preferences on the same pairs: once as inferred from the two other choices from the same trilemma, and once as directly stated after having seen the judgements of the two group members. Comparing socially uninformed and socially informed moral judgements permits a direct comparison of participants' moral judgements on the same moral problem – once made in private, and once after having learned the views of others.

Reward and Punishment

By means of treatments *Reward* and *Punishment* we aim to isolate how participants change their (alleged) moral judgements when they become visible to others and are subject to either favourable or critical evaluation. It is implemented as follows.

Both treatments are in all aspects identical to *Only info*, including the provision of social information. The only exception is an additional phase which comprises peer evaluation of the moral judgements. To this end, all participants take on the role of an evaluator for a different group after all socially informed judgements have been made. Each group is assigned three evaluators, tasked with evaluating the judgements made by the three members of this group. Each group's three evaluators are allocated three distinct sets of twelve dilemmas. Therefore, the respective judgements on all thirty-six solution pairs are evaluated exactly once.

The evaluation goes as follows. An evaluator sees, one after the other, the moral problems (i.e. the pairs of options) together with the three group members' moral judgements on them. By design, for each problem two of the three judgements are socially uninformed while the third is a socially informed one. The evaluator knows this, but not which one is which. Their task is to pick one of the three choices, with the consequence that the person whose choice is picked will receive a reward or a deduction of €10 depending on the treatment (see screenshot B.3 and B.4 in Appendix B.5). At the end of the experiment, one randomly selected evaluation per group is paid out.

Importantly, participants are informed about the fact that their moral judgements will be evaluated and may be subject to rewards (punishments) *after* they have made the first two judgements for each trilemma, and *before* they make their

socially informed choice in light of their group members' answers. Therefore, participants may state a moral preference that they think would maximise (minimise) their chance to be selected for the reward (punishment) given the two judgements of their group members.

Since, along with being asked to make the remaining judgements, participants learn about their group members' moral views as they do in *Only info*, any potential effects on participants' *actual* moral views (caused by the mere provision of information) can be expected to manifest in *Reward* and *Punishment* all the same. To isolate any additional effects in the form of deliberate strategic conformity (or anticonformity) caused by the prospective evaluation, we use *Only info* as a benchmark against which we compare the judgements in *Reward* and *Punishment*.

Moreover, the fact that all participants switch to the role of an evaluator further allows us to examine the relation between evaluation decisions and their own previously made moral judgements.

Further design aspects and discussion

A central design aspect is that we elicit moral judgements, rather than incentivised choices. Since choices can be motivated by many things, with moral considerations being only one of them, eliciting choices would in fact make the data noisier relative to eliciting judgements directly. This constitutes an important distinction to previous studies on moral decision making which use incentivised choices such as Falk and Szech (2013) or Falk et al. (2020). The decisive difference to our study is that they are concerned with how people trade off moral motives against selfish ones. Naturally, this approach renders monetary incentives not only possible, but in fact necessary, as otherwise one would likely mainly elicit cheap talk. Our study instead examines how people trade off

different moral principles, rather than how people trade off moral deeds against material goods.

Incentivising choices would be detrimental to the aim of this study also for a second reason. Moral considerations play a role in a wide area of life. For our experiment, this creates a trade-off between the range (and severity) of the moral problems we include (and hence external validity) and the implementability of participants' choices. In the interest of the former we choose to confront participants with a diverse set of, at times, harsh problems. This implies the elicitation of judgements where any actual implementation is, for obvious reasons, impossible. Examples include extramarital affairs, forced blood transfusions, or selection criteria for the recipients of life-saving organs. The full list of moral problems can be found in Appendix B.2.

Lastly, the description of the trilemmas does not directly address participants. Instead, the trilemmas are presented in an impartial framing and participants are asked what, according to their personal moral judgement, one (the agent) *should* do in such a situation. We use these general and impersonal descriptions, because eliciting what participants *themselves would* do might confound the moral judgement with *a*moral incentives. The impersonal framing avoids this issue by asking what one should do on moral grounds alone. We therefore expect it to more cleanly isolate exclusively moral considerations. In addition, the impersonal wording may mitigate refusals to choose an action due to personal involvement or emotional discomfort, which might otherwise lead to a higher drop-out rate. Moreover, disallowing any solution other than the two proposed may subjectively appear more acceptable if participants are not forced to choose for themselves.

2.2.3 Procedure and Payment

The experiment was conducted in two waves: *Only info*, *Reward*, and *Punishment* were run between the 6th and 15th of November 2021, *No info* on the 16th of November 2022. The entire experimental procedure, including the recruitment, was identical in both waves. We used z-Tree 5.1.3, HROOT, and z-Tree unleashed (Fischbacher, 2007; Bock et al., 2014; Duch et al., 2020). In total, we ran twelve sessions, all online, with a total of $N = 207$ participants (51 in *Only info*, 54 in *Reward*, 48 in *Punishment*, 54 in *No info*). Participants were recruited from the lab subject pool of the University of Konstanz, which explains the relatively small number of participants for an online experiment and the fact that we had no dropouts. All participants were students at the University of Konstanz, on average 23 years old, and 61% of them were female.

Participants in *No info* and *Only info* received a show-up fee of €5. Sessions lasted for approximately one hour, and in both conditions they earned €18.65 on average. The difference between show-up fee and final earnings is due to a fixed amount of money which was paid for each stated moral preference. This was the same for all treatments. In principle, payoffs could have varied, but they did not, as all participants stated a preference on all dilemmas. Since in *Reward* (€+10) and *Punishment* (€-10) one randomly selected evaluation was paid out per group, we adapted the show-up fees such that expected earnings were held constant across treatments. In *Reward*, the show-up fee amounted to €1.50, and in *Punishment* to €8.50. A typical session lasted 1.5 hours in both treatments, for which participants on average earned €22.68 and €23.03 in *Reward* and in *Punishment*, respectively.

Given the potentially sensitive nature of the scenarios, participants could withdraw at any time while retaining the show-up fee. All procedures were approved by the University of Konstanz research ethics committee (Certificate No

20/2021) prior to collecting the data, and we obtained informed consent from all participants.

2.3 Results

Figure 2.1 summarises our main findings. The y-axis depicts the empirical probability of making an intransitive choice. The x-axis indicates whether none, one, or both of the group members stated the opposite opinion. To compare the probabilities in Figure 2.1, we fit linear probability models and perform t-tests for the coefficients β that reflect differences between two probabilities. Cluster robust standard errors account for the dependence among choices made by the same respondent.

The red horizontal reference line at $y = 0.27$ depicts the average intransitivity we observe in *No info*, that is, the baseline intransitivity level *in the absence* of social influence. The causal effect of social influence on moral judgements is graphically represented by the distance of the dots to this reference line. Increases in intransitivity relative to the reference line imply participants change to the other opinion *more* often than in the absence of social influence, decreases imply that participants are more likely to state the predicted opinion.

Conformity in moral judgements The key takeaway from Figure 2.1 is that moral judgements are shaped by social influence. Across all three treatments the relative position of the dots is unambiguously a function of whether their group members share their view or are of different opinion. In other words, depending on what others think, they change their choice with significantly different probabilities. Within all three treatments, the probabilities of intransitivity, represented by the dots, show a pattern consistent with conformity, with the lowest

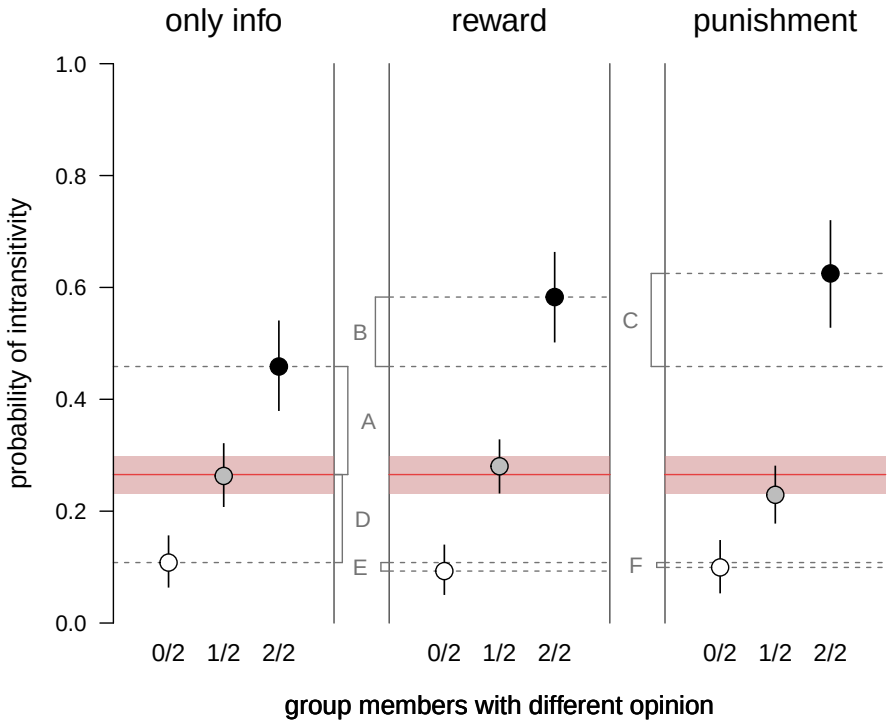


Figure 2.1: Probability of intransitivity across cases and treatments. The dots represent the empirical probability with which the socially informed judgement is unequal to the (predicted) socially uninformed judgement conditional on the number of group members with a different judgement in each treatment. The red horizontal line indicates the baseline intransitivity in the absence of social influence in *No info*. Whiskers and red polygon represent 95% confidence intervals based on a nonparametric block bootstrap (subject ID, 10,000 samples). A: shift of moral preferences; B: strategic shift under reward; C: strategic shift under punishment; D: reinforcement of moral preferences; E: strategic reinforcement under reward; F: strategic reinforcement under punishment

probability of intransitivity in the 0/2 case, followed by 1/2 and 2/2. The differences between the cases 0/2 and 1/2, and those between the cases 1/2 and 2/2 are significant at an α-level of 0.1% in every treatment (t-tests of model coefficients: all $|t| > 3.66$, $df \geq 47$, $p < 0.001$).

Participants conform to the majority opinion, which has two effects. They change their judgement more often upon learning that others hold different views, and they are less likely to change their initial judgement after learning that their group members share their opinion. When social information is mixed, represented by the grey dots, meaning that there is no majority to conform to, they revise their judgements just as often as when no social information is given. Consequently, the intransitivity rate is an upward sloping function of the number of group members who hold the opposite opinion. Although all three treatments consistently produce this pattern of conformity, they permit a different interpretation with regard to the underlying causes. With regard to *Only info*, the results suggest that what people think is morally right is significantly shaped by the opinions of others. While on average participants change to the other option 27% of the time in *No info*, and virtually equally often if the social information is mixed (26.3%), they do so 45.9% of the time if both express a different view, but in only 10.8% of the cases if both group members share the same view. These differences, labelled *A* and *D* in Figure 2.1, indicate intrinsic conformity and are both large in magnitude and significant at an α-level of 0.1% (*A*: $\beta = 0.193$, $t(103) = 4.39$, $p < 0.001$; *D*: $\beta = -0.157$, $t(103) = -5.47$, $p < 0.001$).

Evidence for strategic conformity While in *Only info*, due to the absence of any extrinsic incentives, changes in *stated* views can be interpreted as changes in *actual* views, in *Reward* and *Punishment* such changes may also be strategic in nature. Here, participants are not only informed about the opinions of others,

but their choices are also subject to peer evaluation. Thus, we only observe the joint effect of information and strategic responses to peer evaluation. Together, these effects cause conformity to the majority view. To isolate the mere strategic effect, we compare the intransitivity in *Reward* and *Punishment*, respectively, with that observed in *Only info*. We do this separately for each of the three types of social information (i.e., both group members hold the same, different, or mixed opinions).

In the case of mixed social information, i.e., when one group member holds the same opinion and the other holds the opposite opinion, again we do not see any significant changes, neither relative to the baseline intransitivity, nor to *No info*. This suggests that since there is no majority to which one can conform to, there is, with or without subsequent evaluation, no reason to change to the other option.

However, when participants learn that both of their group members hold different opinions, we find a higher likelihood of them switching to the majority opinion than in *Only info*. These differences are labelled B for positive social incentives in *Reward*, and C for negative social incentives in *Punishment* in Figure 2.1. Both differences are large in magnitude and significant at an α-level of 5% (B: $\beta = 0.124$, $t(103) = 2.14$, $p = 0.035$; C: $\beta = 0.166$, $t(97) = 2.64$, $p = 0.010$). To put this in relation, while intransitivity increases by approximately 20%-points in *Only info*, it increases further by roughly 15%-points in *Reward* and *Punishment*. In sum, under the prospect of peer evaluation, when participants learn that their group members have different views, they switch to the other choice (and thus match their group members') 60% of the time, as opposed to 27%, which we find in the absence of any social influence in *No info*.

If this extra conformity in light of peer evaluation is a general behavioural pattern, we should, in principle, see it too if both group members are of the *same*

opinion. Put differently, if the imminent peer evaluation is a sufficiently strong extra incentive to *switch* to the majority opinion, it should also be a sufficient incentive to *stick* to the majority opinion. Looking at the numbers only, this is indeed what we see: while the probability of intransitivity in *Only info* is 10.8%, it drops to 9.3% in *Reward* and to 10.0% in *Punishment*. These differences in the probability of intransitivity are labelled E and F in Figure 2.1 and are statistically insignificant (E: $\beta = -0.015$, $t(103) = -0.47$, $p = 0.637$; F: $\beta = -0.009$, $t(97) = -0.26$, $p = 0.795$). Because in the *Only info* treatment only about 10% of the participants show intransitivity when both group members are of the same opinion, there is not much room to reduce this intransitivity even further.

Strength of moral preferences matters We also explore the role of the strength of moral conviction for conformity by fitting a linear probability model that controls for the strength of moral preferences. This model is reported in Table B.1 of Appendix B.1 and shows that the strength of preference has a large and significant effect on the probability with which participants change to the other option. Participants are generally less likely to change their moral judgement when their preferences are strong. The effects reported previously are robust when controlling for the strength of moral preferences.

No individual differences in conformity To test for potential individual differences in the response to social influence across participants, we estimate mixture models that vary the number of behavioural types (Dvorak, 2023). Each type is characterised by a distinct vector of three frequencies which describe the frequency of intransitivity, one for each type of social information. We identify the optimal number of behavioural types based on the Bayesian information criterion (Schwarz, 1978). According to the Bayesian information criterion, participants' behaviour in all treatments, *Only info*, *Reward*, as well as in *Pun-*

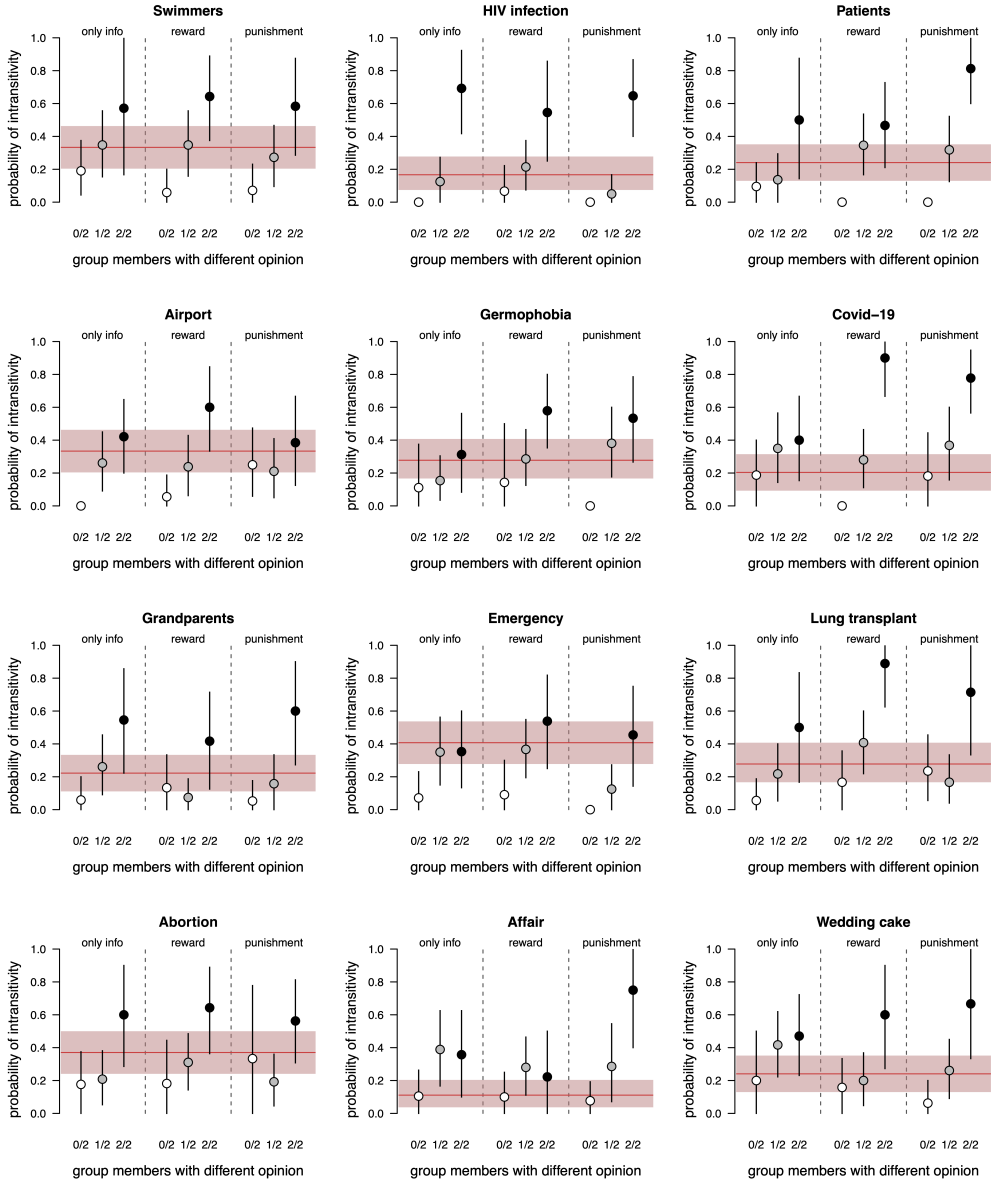


Figure 2.2: Intransitivity across treatments for each moral problem. Dots indicate the frequency of intransitivity. Whiskers represent 95% confidence intervals. The red horizontal line indicates the baseline intransitivity in the absence of social influence in *No info*. Polygons indicate the 95% confidence interval of baseline intransitivity based on nonparametric bootstrapping (subject ID, 10,000 samples).

ishment, is best described by a single behavioural type (Control: one-type BIC = 666, two-types BIC = 679; reward: one-type BIC = 703, two-types BIC = 716; punishment: one-type BIC = 596, two-types BIC = 603). This type is characterised by the average frequencies of intransitivity shown in Figure 2.1. Thus, we find no heterogeneity between participants in the frequency of intransitivity across the three types of social information (0/2, 1/2 and 2/2 in Figure 2.1).

Similar patterns of conformity across moral problems To explore potential differences in conformity between moral problems, Figure 2.2 depicts the treatment-specific levels of intransitivity for each of the 12 moral problems separately. We observe very similar patterns of conformity when looking at the moral problems in isolation. In 30 of 36 treatment-problem combinations (3 treatments $\times$ 12 problems), we observe that the order of the point estimates for the probability of intransitivity is consistent with conformity, with the lowest estimate in the 0/2 case, followed by 1/2 and 2/2. In *Only info*, the point estimates for the probability of intransitivity when both group members indicated a different opinion (case 2/2) lie above the baseline intransitivity for 11 out of 12 dilemmas, which suggests a shift in moral judgements in the absence of strategic considerations. In *Reward* (*Punishment*) 8 (9) out of 12 point estimates for the same case lie above the point estimate shown for *Only info*, which suggests that peer evaluation additionally increases conformity.

Strategic incentives for conformity The finding that participants conform even more when facing peer evaluation raises the question whether this behaviour is economically justified. In other words, does conformity maximise participants' chances of being selected by the evaluator in *Reward*, and minimise it in *Punishment*? To answer this question, we need to look at the probability with which evaluators select a unique answer - as opposed to one from

the majority. If all players choose the same option, the selection probability equals $1/3$. Whether the selection probability of a unique answer is higher or lower than $1/3$ determines whether it is better to choose the unique option or to choose the same option as the other two players. If this probability is higher than $1/3$, going with the others would be optimal under punishment and going alone would be optimal under reward.

The probability of the unique choice being selected is 0.45 in *Reward* and 0.51 in *Punishment*, and thus significantly different from $1/3$ in both treatments (*Reward*: $\beta = 0.154$, $t(53) = 5.45$, $p < 0.001$; *Punishment*: $\beta = 0.176$, $t(47) = 4.71$, $p < 0.001$). This means that the conformity to the majority view we observe in both treatments does pay off economically for participants in *Punishment*, but lowers expected payoffs in *Reward*. In *Punishment* participants will successfully 'hide in the crowd' and thereby avoid a payoff deduction roughly two-thirds of the time, while anticonformity would imply being chosen roughly $1/2$ of the time. In *Reward* participants forgo expected payoffs from rewards due to their tendency to conform to the views of others. The fact that the unique answer is selected with probability 0.45 implies that by choosing a unique answer they could increase their chances of receiving the reward by about 12 percentage points compared to when they conform with their group members.

Homophily explains evaluation behaviour What motivates evaluators to reward participant or punish another? To explore this, we run a linear probability model on evaluation behaviour for both *Reward* and *Punishment*. We focus on situations where both alternatives are chosen. This allows us to study whether evaluators reward or punish unique choices (salience) or according to their own moral preference (homophily). In this model the dependent variable is a dummy which indicates whether a choice is selected for reward/punishment. It is re-

	Treatment	
	Reward	**Punishment**
Unique choice	-0.01	-0.01
	(0.04)	(0.06)
Own preference	0.81***	-0.57***
	(0.04)	(0.06)
Constant	0.08***	0.81***
	(0.03)	(0.06)
Observations	271	204
N	54	48

Table 2.1: Linear probability models for evaluation behaviour. Table shows coefficients of linear probability models with cluster robust standard errors in parentheses. The dependent variable takes on the value 1 if the choice is selected for reward/punishment, and 0 otherwise. *Unique choice* is a dummy variable which indicates whether the choice was a unique choice relative those of the group members, *Own preference* is a dummy indicating whether the choice matched the evaluator's preference. *** $p < 0.01$.

gressed on the dummy variables *Unique choice* and *Own preference* which indicate whether a socially informed judgement was unique relative to the two group members' and whether it matched the evaluator's moral preference respectively. The results are shown in Table 2.1. They suggest that homophily clearly drives evaluators' decisions. While minority choices are not selected more or less often just because they are unique judgements, choices in line with the evaluator's judgement are 81% more likely to be rewarded and 57% less likely to be punished (Reward: $\beta = 0.81$, $t(53) = 20.85$, $p < 0.001$, Punishment: $\beta = -0.57$, $t(47) = -9.49$, $p < 0.001$).

2.4 Conclusion and Discussion

How and to what extent moral beliefs are shaped by social influence has been a long-standing question in the social sciences. Here, we report the results of an online experiment that provides strong evidence for conformity to the moral judgements of others. Our results document that conformity is caused by genuine changes in moral views, and further amplified by strategic distortions under endogenously assigned rewards and punishments. The effect sizes are statistically significant and large in magnitude, and persist when controlling for the strength of moral conviction towards which social influence is exerted. Notably, conformity in our experiment extends to situations where participants are exposed to reward mechanisms. This contrasts sharply with theoretical predictions and empirical evidence from *a*moral domains (Dvorak et al., 2025), and supports the notion that human decision making in moral domains cannot be adequately explained by preferences over monetary payoffs alone. With regard to the motives behind evaluation behaviour, that is, the selection of moral choices for reward and punishment, we find strong evidence for homophily, with evaluators predominantly rewarding their own and punishing deviant moral judgements.

Our study suggests that political controversies related to the moral sphere are substantially shaped by the opinions of others that people encounter, which seems especially relevant in the age of social media. The observed patterns illuminate the mechanisms underlying the fragmentation of moral views in society, which are often at the heart of deeply entrenched political controversies and add to the growing disharmony, mutual intolerance, and polarisation that many societies are currently experiencing (Converse, 2006; Baldassarri and Gelman, 2008; Schweighofer et al., 2020), and which are widely debated both within and outside academia (Abramowitz and Saunders, 2008; Hare and Poole, 2014; Latané and Bourgeois, 2001).

General Conclusion

This thesis presents novel evidence of social influence. It thus contributes to a large literature spanning the fields of economics and the social sciences more generally, on a topic that has plausibly attracted human curiosity for much longer than any of these disciplines have been formally established. Chapter 1 addresses the question *'Are peer effects shaped by social closeness?'*, and Chapter 2 asks '*Are moral judgements prone to social influence?'*. In the following, the key takeaways of both chapters are summarised.

The main finding from Chapter 1 shows that peer effects on cooperation are contingent on social closeness. Our results suggest that predictions about how a peer will affect a decision maker's cooperativeness become significantly more accurate once the decision maker's social closeness to that peer is taken into account. Crucially, our results further suggest that the impact of social closeness on the predicted peer influence hinges on the peer's role in the interaction: if the peer is a passive observer, greater social closeness tends to dampen the effects, causing a lower level of cooperation. In contrast, if the peer makes their own choice, social closeness amplifies the influence, causing decision makers to adopt a similar level of cooperativeness.

With respect to underlying mechanisms, the results highlight the importance of belief-based channels in strategic settings, but crucially also show that these are insensitive to social closeness. Rather, the moderating effect of social closeness

on peer influence arises through direct effects on behaviour.

To examine this research question, we conducted a controlled laboratory experiment in which participants made a one-shot contribution decision to a public good. Our novel design simultaneously (i) randomly assigned individuals to the role of an observing or an observed player, and (ii) exogenously varied the degree of social closeness between the two. The experiment was designed to rule out competing channels of influence: peer effects cannot be explained by direct payoff externalities, by social-preference links between peers, or by any form of learning about how the game operates. We further controlled for *ex ante* existing cooperation preferences and expectations of co-players' choices.

The key finding from Chapter 2 is that people exhibit a marked tendency to conform to the moral judgements of others. Importantly, this conformity manifests both as genuine opinion change, where individuals adopt the views of others, and as strategic misrepresentation, where individuals adjust their stated opinions toward the majority position in order to attract social reward or avoid sanctions. This overall pattern is representative of the individual moral problems used in our experiment. We also examined the motives guiding evaluators in assigning rewards or sanctions to publicly stated moral judgements. The results show clearly that evaluators reward statements consistent with their own privately held views and punish those that conflict with them.

The findings reported in Chapter 2 are based on an online experiment in which participants were confronted with a series of moral dilemmas. Across treatments, we systematically varied the form of social influence to which participants were exposed when expressing their judgements on these dilemmas. In one condition, participants received only information about the opinions of others. In two others, they were additionally subject to endogenous social evaluation, facing the possibility of rewards or sanctions for their stated judgements.

To isolate the role of social influence from other sources of moral opinion change, we included a control condition in which participants formed judgements in the absence of any social influence.

Taken together, Chapters 1 and 2 underscore the central role of social influence in shaping both values and choices. Although the two chapters differ in important respects - for instance, one investigates cooperation in a public good game, the other moral judgement; one relies on exogenous monetary incentives, the other on endogenous evaluation or no incentives at all; one focuses on how agents influence one another through observable choices, the other on conformity in moral beliefs - these contrasts reinforce rather than weaken the broader lesson. The key takeaway from this thesis is that social influence shapes outcomes not only in exceptional cases but across a wide range of contexts, affecting both what we think and how we act, often in response to surprisingly minimal social cues.

These findings, however, also raise unresolved questions and highlight promising directions for future research. Most importantly, they call for theoretical clarification. Despite the breadth of the literature on social preferences, existing models cannot readily explain the causal role of social closeness in peer effects documented in this thesis, nor do they explain the seemingly *ad hoc*, genuine shifts in moral opinions that arise in response to astonishingly little social information.

This thesis opened with Aristotle, and hence it may be fitting to close on a philosophical note and a final takeaway. We found that even when nothing seems to be at stake, subtle social cues can matter to others, implying that our choices carry a quiet responsibility.

Bibliography

Abramowitz, A. I. and Saunders, K. L. (2008). Is polarization a myth? *The Journal of Politics*, 70(2):542–555.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The Quarterly Journal of Economics*, 94(4):749–775.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, 97(4):543–569.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, 95(9-10):1082–1095.

Allen, V. L. (1965). Situational factors in conformity. In *Advances in Experimental Social Psychology*, volume 2, pages 133–175. Elsevier.

Amabile, T. M., Goldfarb, P., and Brackfleld, S. C. (1990). Social influences on creativity: Evaluation, coaction, and surveillance. *Creativity Research Journal*, 3(1):6–21.

Anderson, L. R. and Holt, C. A. (1997). Information cascades in the laboratory. *The American Economic Review*, pages 847–862.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477.

Apesteguia, J., Huck, S., and Oechssler, J. (2007). Imitation—theory and experimental evidence. *Journal of Economic Theory*, 136(1):217–235.

Arendt, H. (1998). *The Human Condition*, chapter II.7, page 50. University of Chicago Press, Chicago, 2 edition. Original work published 1958.

Aristotle (1998). *Politics*. Hackett, Indianapolis. Original work published ca. 350 BCE.

Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4):596.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, Leadership, and Men*, pages 222–236.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, 193(5):31–35.

Asch, S. E. (1956). Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological Monographs: General and applied*, 70(9):1.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729):59–64.

Ayres, I., Raseman, S., and Shih, A. (2013). Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. *The Journal of Law, Economics, & Organization*, 29(5):992–1022.

Baldassarri, D. and Gelman, A. (2008). Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817.

Bateson, M., Nettle, D., and Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, 2(3):412–414.

Bayer, P., Hjalmarsson, R., and Pozen, D. (2009). Building criminal capital behind bars: Peer effects in juvenile corrections. *The Quarterly Journal of Economics*, 124(1):105–147.

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy*, 82(6):1063–1093.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *The American Economic Review*, 96(5):1652–1678.

Benjamin, D. J., Choi, J. J., and Fisher, G. (2016). Religious identity and economic behavior. *Review of Economics and Statistics*, 98(4):617–637.

Benjamin, D. J., Choi, J. J., and Strickland, A. J. (2010). Social identity and preferences. *The American Economic Review*, 100(4):1913–1928.

Bernedo, M., Ferraro, P. J., and Price, M. (2014). The persistent impacts of norm-based messaging and their implications for water conservation. *Journal of Consumer Policy*, 37(3):437–452.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, 102(5):841–877.

Berscheid, E., Snyder, M., and Omoto, A. M. (1989). The relationship closeness inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology*, 57(5):792.

Bertrand, J. (1883). Théorie mathématique de la richesse sociale [par léon walras]; recherches sur les principes mathématiques de la théorie des richesses [par augustin cournot]. *Journal des Savants*, September:499–508.

Beshears, J., Choi, J. J., Laibson, D., Madrian, B. C., and Milkman, K. L. (2015). The effect of providing peer information on retirement savings decisions. *The Journal of Finance*, 70(3):1161–1201.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives*, 12(3):151–170.

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.

Bollinger, B. and Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, 31(6):900–912.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *The American Economic Review*, 91(1):166–193.

Bostyn, D. H. and Roets, A. (2017). An asymmetric moral conformity effect: Subjects conform to deontological but not consequentialist majorities. *Social Psychological and Personality Science*, 8(3):323–330.

Bourdieu, P. (1977). *Outline of a Theory of Practice*. Cambridge University Press, Cambridge. Translated by Richard Nice. First published 1972 entitled *Esquisse d'une théorie de la pratique*, Libraire Dros S. A., Switzerland.

Bowles, S. and Choi, J.-K. (2013). Coevolution of farming and private property during the early holocene. *Proceedings of the National Academy of Sciences*, 110(22):8830–8835.

Bowles, S. and Gintis, H. (2003). Origins of human cooperation. *Genetic and Cultural Evolution of Cooperation*, 2003:429–43.

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535.

Boyd, R. and Richerson, P. J. (1988). *Culture and the Evolutionary Process*. University of Chicago Press.

Bradley, A., Lawrence, C., and Ferguson, E. (2018). Does observability affect prosociality? *Proceedings of the Royal Society B: Biological Sciences*, 285(1875):20180116.

Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, 82(4):1273–1301.

Bursztyn, L., Egorov, G., and Jensen, R. (2019a). Cool to be smart or smart to be cool? Understanding peer pressure in education. *The Review of Economic Studies*, 86(4):1487–1526.

Bursztyn, L., Fiorin, S., Gottlieb, D., and Kanz, M. (2019b). Moral incentives in credit card debt repayment: evidence from a field experiment. *Journal of Political Economy*, 127(4):1641–1683.

Bursztyn, L. and Jensen, R. (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics*, 130(3):1329–1367.

Charness, G. and Chen, Y. (2020). Social identity, group behavior, and teams. *Annual Review of Economics*, 12(1):691–713.

Chaudhuri, A., Paichayontvijit, T., et al. (2006). Conditional cooperation and voluntary contributions to a public good. *Economics Bulletin*, 3(8):1–14.

Cialdini, R. B. (1984). *Influence: How and Why People Agree to Things*. William Morrow, New York.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621.

Cicero, M. T. (1927). *Tusculan Disputations*. Loeb Classical Library 141. Loeb Classical Library. Harvard University Press, Cambridge, MA. Original work published 45 BCE.

Converse, P. E. (2006). The nature of belief systems in mass publics (1964). *Critical Review*, 18(1-3):1–74.

Cournot, A.-A. (1897). *Researches into the Mathematical Principles of the Theory of Wealth*. The Macmillan Company; Macmillan, New York and London. With a bibliography of mathematical economics by Irving Fisher; English translation of *Recherches sur les principes mathématiques de la théorie des richesses* (1838), L. Hachette.

Crawford, V. and Sobel, P. J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Crosetto, P., Weisel, O., and Winter, F. (2012). A flexible z-Tree implementation of the social value orientation slider measure (Murphy et al. 2011)–manual. *Jena Economic Research Paper*, 62.

Curry, O. S., Mullins, D. A., and Whitehouse, H. (2019). Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, 60(1):47–69.

Dahl, G. B., Løken, K. V., and Mogstad, M. (2014). Peer effects in program participation. *The American Economic Review*, 104(7):2049–74.

de Bornier, J. M. (1992). The 'cournot-bertrand debate': A historical perspective. *History of Political Economy*, 24(3):623–656.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, 127(1):1–56.

Denton, K. K., Ram, Y., Liberman, U., and Feldman, M. W. (2020). Cultural evolution of conformity and anticonformity. *Proceedings of the National Academy of Sciences*, 117(24):13603–13614.

Duch, M. L., Grossmann, M. R., and Lauer, T. (2020). z-tree unleashed: A novel client-integrating architecture for conducting z-tree experiments over the internet. *Journal of Behavioral and Experimental Finance*, 28:100400.

Duflo, E. and Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly Journal of Economics*, 118(3):815–842.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1):197–228.

Durkheim, É. (1938). *The Rules of Sociological Method*. The Free Press, New York; Collier-Macmillan Limited, London, New York; London, 8 edition.

Translated by Sarah A. Solovay and John H. Mueller. Original work published 1895.

Dvorak, F. (2023). Stratest: A software package for strategy frequency estimation. *Journal of the Economic Science Association*, 9(2):337–349.

Dvorak, F., Fischbacher, U., and Schmelz, K. (2025). Strategic conformity or anti-conformity to avoid punishment and attract reward. *The Economic Journal*, 135(666):556–583.

Edgeworth, F. Y. (1925). The pure theory of monopoly. In Society, R. E., editor, *Papers Relating to Political Economy*, volume 1, pages 111–142. Macmillan, London. English translation of *La teoria pura del monopolio* (1897). Source: *Bibliothèque nationale de France's Gallica* website.

Efferson, C., Bernhard, H., Fischbacher, U., and Fehr, E. (2024). Super-additive cooperation. *Nature*, 626(8001):1034–1041.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117.

Enke, B. (2019). Kinship, cooperation, and the evolution of moral systems. *The Quarterly Journal of Economics*, 134(2):953–1019.

Enke, B. (2020a). Moral values and voting. *Journal of Political Economy*, 128(10):3679–3729.

Enke, B. (2020b). What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398.

Enke, B., Rodríguez-Padilla, R., and Zimmermann, F. (2023). Moral universalism and the structure of ideology. *The Review of Economic Studies*, 90(4):1934–1962.

Falk, A. and Ichino, A. (2006). Clean evidence on peer effects. *Journal of Labor Economics*, 24(1):39–57.

Falk, A., Neuber, T., and Szech, N. (2020). Diffusion of being pivotal and immoral outcomes. *The Review of Economic Studies*, 87(5):2205–2229.

Falk, A. and Szech, N. (2013). Morals and markets. *Science*, 340(6133):707–711.

Falk, A. and Zimmermann, F. (2017). Consistency as a signal of skills. *Management Science*, 63(7):2197–2210.

Feather, N. T. (1995). Values, valences, and choice: The influences of values on the perceived attractiveness and choice of alternatives. *Journal of Personality and Social Psychology*, 68(6):1135.

Fehr, E. (2004). Don't lose your reputation. *Nature*, 432(7016):449–450.

Fehr, E. and Charness, G. (2025). Social preferences: fundamental characteristics and economic consequences. *Journal of Economic Literature*, 63(2):440–514.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785–791.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4):980–994.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Économics*, 114(3):817–868.

Fehrler, S. and Hughes, N. (2018). How transparency kills information aggregation: theory and experiment. *American Economic Journal: Microeconomics*, 10(1):181–209.

Ferraro, P. J., Miranda, J. J., and Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *The American Economic Review*, 101(3):318–322.

Ferraro, P. J. and Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1):64–73.

Festinger, L. (1953). An analysis of compliant behavior. In Sherif, M. and Wilson, M. O., editors, *Group Relations at the Crossroads*, pages 232–256. Harper, New York.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Fischbacher, U. and Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *The American Economic Review*, 100(1):541–56.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5.

Foucault, M. (1980). *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977*. Pantheon Books, New York.

Funk, P. (2010). Social incentives and voter turnout: Evidence from the Swiss mail ballot system. *Journal of the European Economic Association*, 8(5):1077–1103.

Gächter, S. and Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1518):791–806.

Gächter, S., Kölle, F., and Quercia, S. (2017). Reciprocity and the tragedies of maintaining and providing the commons. *Nature Human Behaviour*, 1(9):650–656.

Gächter, S., Kölle, F., and Quercia, S. (2022). Preferences and perceptions in provision and maintenance public goods. *Games and Economic Behavior*, 135:338–355.

Gächter, S., Nosenzo, D., and Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences? *Journal of the European Economic Association*, 11(3):548–573.

Gächter, S., Starmer, C., and Tufano, F. (2015). Measuring the closeness of relationships: A comprehensive evaluation of the 'inclusion of the other in the self' scale. *PloS One*, 10(6):e0129478.

Gächter, S., Starmer, C., and Tufano, F. (2023). Measuring "group cohesion" to reveal the power of social relationships in team production. *Review of Economics and Statistics*, pages 1–45.

Gerber, A. S., Green, D. P., and Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1):33–48.

Gintis, H., Henrich, J., Bowles, S., Boyd, R., and Fehr, E. (2008). Strong reciprocity and the roots of human morality. *Social Justice Research*, 21(2):241–253.

Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3):481–510.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Grimalda, G., Pondorfer, A., and Tracer, D. P. (2016). Social image concerns promote cooperation more than altruistic punishment. *Nature Communications*, 7(1):1–8.

Grose, R. G., Hayford, S. R., Cheong, Y. F., Garver, S., Kandala, N. B., and Yount, K. M. (2019). Community influences on female genital mutilation/-cutting in kenya: Norms, opportunities, and ethnic diversity. *Journal of Health and Social Behavior*, 60(1):84–100.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408.

Habermas, J. (1987). *The Theory of Communicative Action, Volume 2: Lifeworld and System - A Critique of Functionalist Reason*. Polity Press, Cambridge. First published in German as *Theorie des kommunikativen Handelns*, Bd. 2 (Suhrkamp, 1981).

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316(5827):998–1002.

Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage.

Hare, C. and Poole, K. T. (2014). The polarization of contemporary american politics. *Polity*, 46(3):411–429.

Hayek, F. A. (1945). The use of knowledge in society. *The American Economic Review*, 35(4):519–530.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., et al. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 327(5972):1480–1484.

Herbst, D. and Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science*, 350(6260):545–549.

Hitlin, S. and Piliavin, J. A. (2004). Values: Reviving a dormant concept. *Annual Review of Sociology*, 30(1):359–393.

Hu, L.-t. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55.

Hume, D. (1975). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Clarendon Press, Oxford. First published 1748 (EHU), 1751 (EPM).

Hume, D. (1978). *A Treatise of Human Nature*. Clarendon Press, Oxford. First published 1739–1740.

Isler, O., Gächter, S., Maule, A. J., and Starmer, C. (2021). Contextualised strong reciprocity explains selfless cooperation despite selfish intuitions and weak social heuristics. *Scientific Reports*, 11(1):1–17.

Kelman, H. C. (1961). Processes of opinion change. *Public Opinion Quarterly*, 25(1):57–78.

Kirchler, M., Huber, J., Stefan, M., and Sutter, M. (2016). Market design and moral behavior. *Management Science*, 62(9):2615–2625.

Kreps, D. M. and Wilson, R. (1982). Reputation and imperfect information. *Journal of Economic Theory*, 27(2):253–279.

Kundu, P. and Cummins, D. D. (2013). Morality and conformity: The asch paradigm applied to moral decisions. *Social Influence*, 8(4):268–279.

Kuran, T. (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press, Cambridge, MA.

Latané, B. and Bourgeois, M. J. (2001). Dynamic social impact and the consolidation, clustering, correlation, and continuing diversity of culture. *Blackwell Handbook of Social Psychology: Group Processes*, pages 235–258.

Locke, J. (1975). *An Essay Concerning Human Understanding*. Clarendon Press, Oxford. First published 1690.

Mas, A. and Moretti, E. (2009). Peers at work. *The American Economic Review*, 99(1):112–45.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8):771–781.

Myerson, R. B. (1999). Nash equilibrium and the history of economic theory. *Journal of Economic Literature*, 37(3):1067–1082.

Nash Jr, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49.

Nash Jr, J. F. (1951). Non-cooperative games. *Annals of Mathematics*, 54(2):286–295.

Parsons, T. (1949). *The Structure of Social Action*, volume 491. Free press New York. Original publication 1937.

Parsons, T. and Shils, E. A. (1951). *Toward a General Theory of Action*. Harvard University Press, Cambridge, MA and London, England.

Plato (1994). *Gorgias*. Oxford World's Classics. Oxford University Press, Oxford. Original work published ca. 380 BCE.

Pruckner, G. J. and Sausgruber, R. (2013). Honesty on the streets: A field study on newspaper purchasing. *Journal of the European Economic Association*, 11(3):661–679.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.

Rand, D. G. and Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, 17(8):413–425.

Rege, M. and Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics*, 88(7-8):1625–1644.

Rohan, M. J. (2000). A rose by any name? the values construct. *Personality and Social Psychology Review*, 4(3):255–277.

Rom, S. C. and Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, 74:24–37.

Rousseau, J.-J. (1964). Discours sur l'origine et les fondements de l'inégalité parmi les hommes. In Gagnebin, B. and Raymond, M., editors, *Œuvres*

complètes, volume 3 of *Bibliothèque de la Pléiade*. Gallimard, Paris. Original work published 1755.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704.

Sato, R. and Takasaki, Y. (2019). Peer effects on vaccination behavior: experimental evidence from rural nigeria. *Economic Development and Cultural Change*, 68(1):93–129.

Schelling, T. C. (2015). *The Strategy of Conflict*. Pickle Partners Publishing. Reprint of the 1960 edition.

Schlag, K. H. (1998). Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory*, 78(1):130–156.

Schlag, K. H. (1999). Which one should I imitate? *Journal of Mathematical Economics*, 31(4):493–522.

Schwartz, S. H. (2012). An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464.

Schweighofer, S., Schweitzer, F., and Garcia, D. (2020). A weighted balance model of opinion hyperpolarization. *Journal of Artificial Societies and Social Simulation*, 23(3):5.

Shalley, C. E. and Perry-Smith, J. E. (2001). Effects of social-psychological factors on creative performance: The role of informational and controlling

expected evaluation and modeling experience. *Organizational Behavior and Human Decision Processes*, 84(1):1–22.

Shayo, M. (2020). Social identity and economic policy. *Annual Review of Economics*, 12(1):355–389.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology (Columbia University)*.

Sherif, M. (1936). *The Psychology of Social Norms.* Harper.

Soll, J. B. and Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):780.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3):335–374.

Tetlock, P. E. (1986). A value pluralism model of ideological reasoning. *Journal of Personality and Social Psychology*, 50(4):819.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 59(2):204–217.

Thöni, C. and Gächter, S. (2015). Peer effects and social preferences in voluntary cooperation: A theoretical and experimental analysis. *Journal of Economic Psychology*, 48:72–88.

Thöni, C. and Volk, S. (2018). Conditional cooperation: Review and refinement. *Economics Letters*, 171:37–40.

Tomasello, M. (2016). *A Natural History of Human Morality*. Harvard University Press.

Veblen, T. (1922). *The Theory of the Leisure Class: An Economic Study of Institutions*. B. W. Huebsch, New York. New edition (Huebsch); originally copyrighted 1899 and 1912 by The Macmillan Company.

Vega-Redondo, F. (1997). The evolution of walrasian behavior. *Econometrica*, 65(2):375–384.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1st edition.

Weber, M. (1978). *Economy and Society: An Outline of Interpretive Sociology*. University of California Press, Berkeley and Los Angeles, California; University of California Press, Ldt., London. 2 vols.; Reissue of the 1968 ed.; based on the 4th German ed. entitled *Wirtschaft und Gesellschaft – Grundriß der verstehenden Soziologie* of 1956.

Welch, I. (1992). Sequential sales, learning, and cascades. *The Journal of Finance*, 47(2):695–732.

Yoeli, E., Hoffman, M., Rand, D. G., and Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences*, 110(supplement_2):10424–10429.

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, 7(1):359–387.

Appendices

Appendix A

Chapter 1

A.1 Supplementary Figures

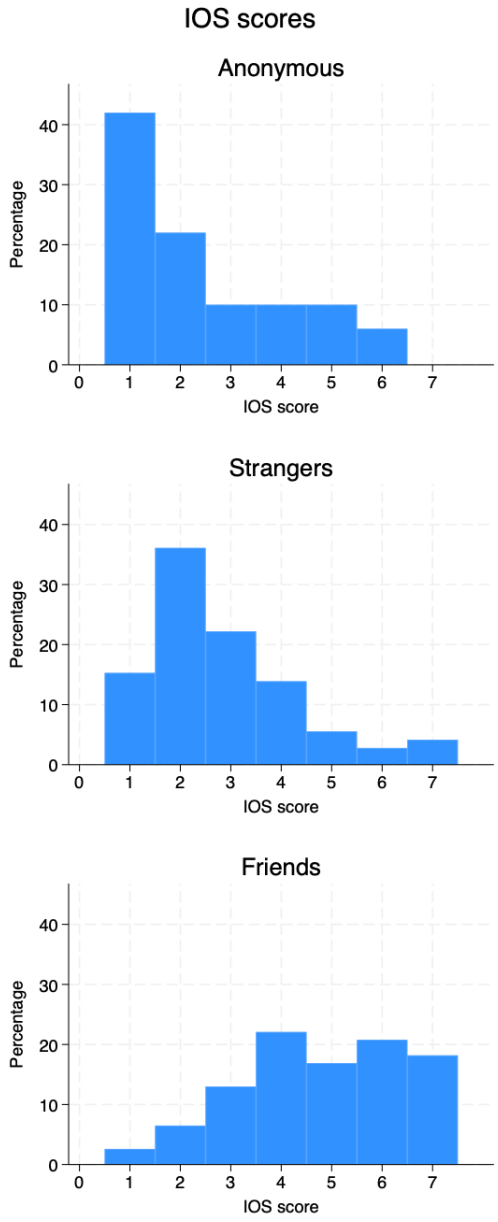


Figure A.1: Histograms plot the distribution of IOS scores of interacting peers in each treatment. IOS scores increase from treatment *Anonymous* to *Stranger* to *Friend* ($p < 0.001$), underscoring that the treatment manipulation is psychologically meaningful.

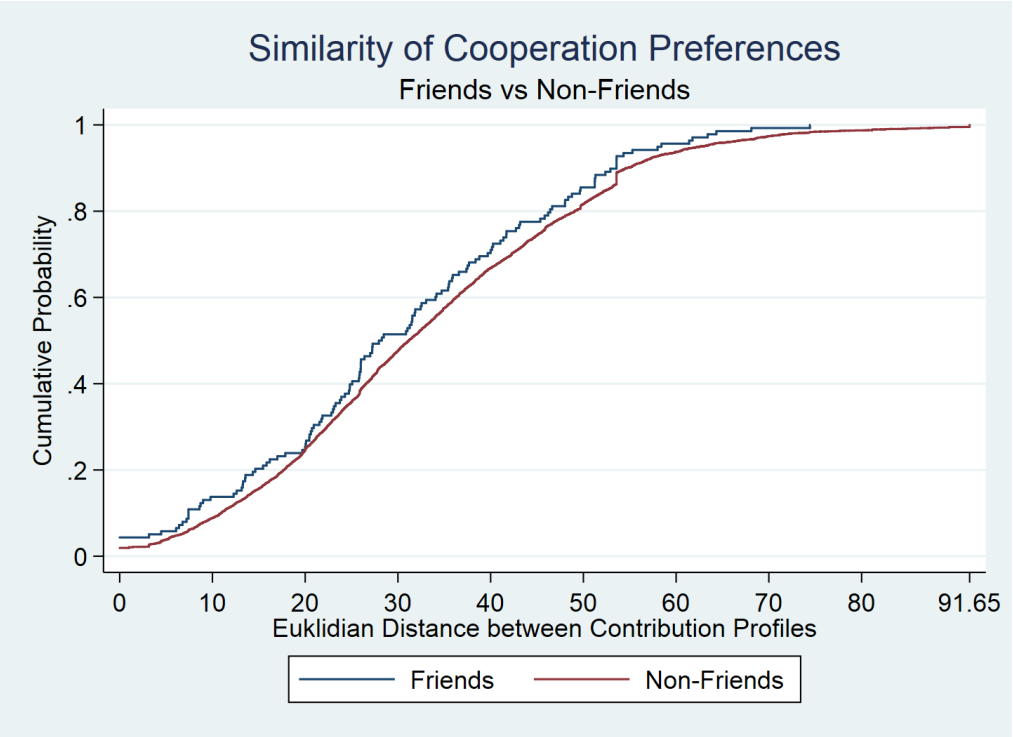


Figure A.2: Graph plots CDFs of Euclidean Distances between contribution preferences as elicited in Phase 1, the strategy method public good game, of our experiment. Non-parametric analysis indicates that friends have more similar contribution preferences (p = 0.02 , Mann-Whitney U test, 1-sided).

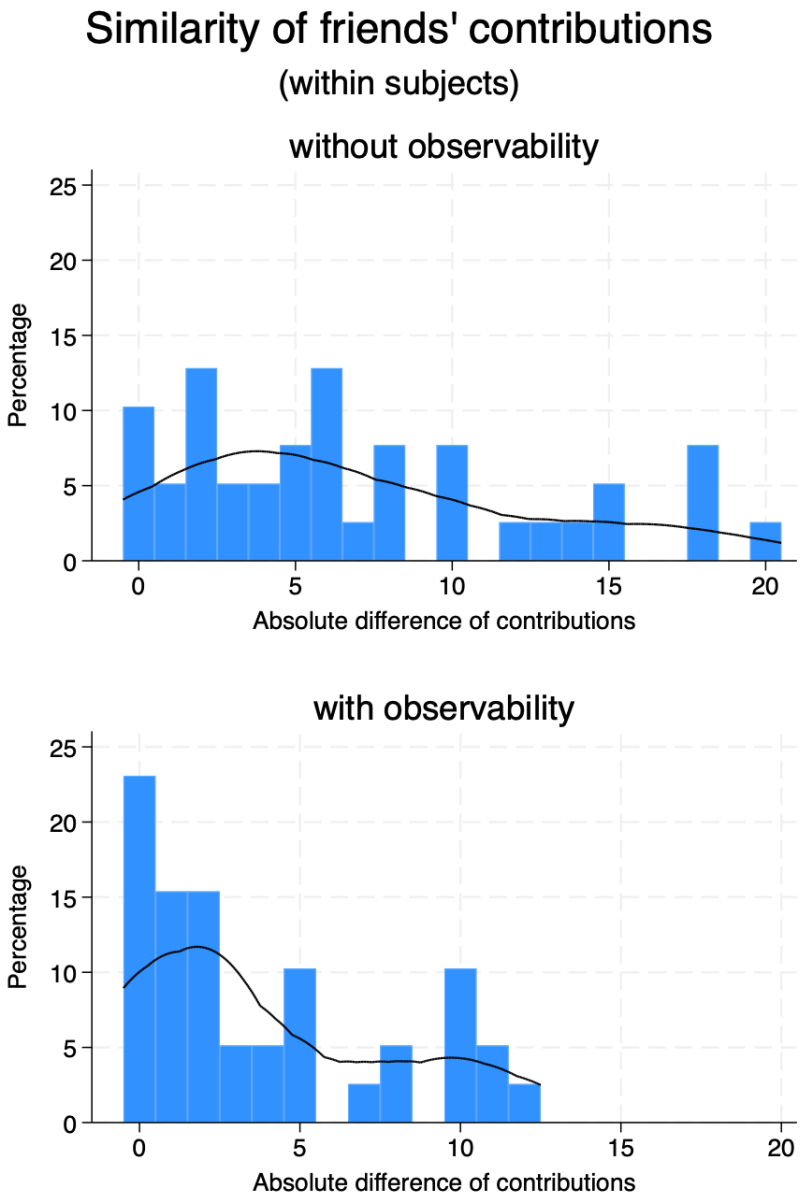


Figure A.3: Histograms show the absolute differences of contribution choices of pairs of friends in Phase 1 (top) and Phase 2 (bottom). A Wilcoxon signed-rank test rejects the null of equal distributions ($p = 0.001$), indicating that friends' contribution choices are more similar when one observes the other than in the absence of observability.

A.2 Supplementary Tables

Table A.1: Structural Equation Model Estimates – Observed Participants

	Estimated Equation		
Predictor Variable	Belief about Avg. Contribution	ABC Prediction	Contribution
Treatment: Anonymous	−0.084		2.691***
	(1.077)		(0.762)
Treatment: Stranger	0.951		1.157
	(1.027)		(0.790)
Treatment: Friend	0.161		0.551
	(1.146)		(0.935)
Belief about Avg. Contribution		0.801***	0.479***
		(0.092)	(0.101)
ABC Prediction			0.474***
			(0.082)
Constant	10.455***	0.686	−0.549
	(0.715)	(0.920)	(0.574)
R^2	0.005	0.362	0.634
Observations		169	

Notes: The three columns represent simultaneously estimated structural equations. The sample is restricted to observed participants and participants in *No Observation*. The variables representing beliefs and ABC predictions are endogenously determined within the model and serve as mediators of treatment effects on contributions. Estimates were obtained using Huber-White robust standard errors (in parentheses) to account for potential heteroskedasticity and non-normality in residuals. R^2 values are reported for each endogenous equation. Likelihood-based fit statistics are not reported due to robust estimation. Residual-based fit indices: SRMR = 0.040, CD = 0.056. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.2: Decomposition of Effects by Treatment - Observed

Effect Type	**Treatment**	**Estimate**	**SE**	**p-value**	**95% CI**
Direct Effect	Anonymous	2.691	0.762	0.000	[1.197, 4.186]
	Stranger	1.157	0.790	0.143	[−0.391, 2.705]
	Friend	0.551	0.935	0.556	[−1.282, 2.383]
Indirect Effect	Anonymous	−0.072	0.924	0.938	[−1.883, 1.739]
	Stranger	0.817	0.897	0.363	[−0.942, 2.575]
	Friend	0.138	0.983	0.888	[−1.788, 2.065]
Total Effect	Anonymous	2.619	1.226	0.033	[0.216, 5.022]
	Stranger	1.973	1.239	0.111	[−0.454, 4.401]
	Friend	0.689	1.308	0.599	[−1.875, 3.253]

Notes: Estimates indicate effects of being observed by a peer and are based on structural equation modelling using Huber-White robust standard errors. Indirect effects are the combined effects transmitted through beliefs and ABC predictions. Total effects are calculated as the sum of direct and (combined) indirect effects.

Table A.3: Structural Equation Model Estimates – Observing Participants

	Estimated Equation		
Predictor Variable	Belief about Avg. Contribution	ABC Prediction	Contribution
Observed Contribution	0.253*		−0.201
	(0.135)		(0.178)
Treatment: Stranger	−3.144		−1.653
	(1.988)		(2.857)
Treatment: Friend	−3.464*		−3.512
	(1.953)		(2.877)
Observed Contribution × Treatment: Stranger	0.185		0.218
	(0.166)		(0.188)
Observed Contribution × Treatment: Friend	0.115		0.413**
	(0.169)		(0.191)
Belief about Avg. Contribution		1.001***	0.380**
		(0.068)	(0.191)
ABC Prediction			0.566***
			(0.098)
Constant	9.223***	−1.122	3.186
	(1.560)	(0.717)	(2.766)
R^2	0.358	0.527	0.645
Observations	103		

Notes: The three columns represent simultaneously estimated structural equations. The sample is restricted to observing participants and participants in treatment *Anonymous*, *Stranger* and *Friend*. The variables representing beliefs and ABC predictions are endogenously determined within the model and serve as mediators of treatment effects on contributions. Estimates were obtained using Huber-White robust standard errors (in parentheses) to account for potential heteroskedasticity and non-normality in residuals. R^2 values are reported for each endogenous equation. Likelihood-based fit statistics are not reported due to robust estimation. Residual-based fit indices: SRMR = 0.061, CD = 0.408. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.4: Decomposition of Effects by Treatment - Observers

Effect Type	Treatment	Estimate	SE	p-value	95% CI
Direct Effect	Anonymous	−0.201	0.178	0.259	[−0.549, 0.148]
	Stranger	0.017	0.120	0.886	[−0.218, 0.252]
	Friend	0.212	0.140	0.129	[−0.062, 0.486]
Indirect Effect	Anonymous	0.239	0.128	0.061	[−0.011, 0.490]
	Stranger	0.414	0.113	0.000	[0.193, 0.636]
	Friend	0.348	0.127	0.006	[0.100, 0.597]
Total Effect	Anonymous	0.039	0.250	0.877	[−0.450, 0.528]
	Stranger	0.432	0.137	0.002	[0.162, 0.701]
	Friend	0.560	0.092	0.000	[0.380, 0.740]

Notes: Estimates indicate observing participants' responsiveness to a unit increase in the contribution of their peer and are based on structural equation modelling using Huber-White robust standard errors. Indirect effects are the combined effects transmitted through beliefs and ABC predictions. Total effects are calculated as the sum of direct and (combined) indirect effects.

A.3 Instructions

Below are English translations of the instruction sheets handed out to participants in the course of the experiment. Each sheet was handed out in all treatments. Treatment differences occur only on Sheet 3, which corresponds to Part 2 of the experiment, and are highlighted in red.

[Translated from German]

Sheet 1 - Welcome Sheet

Welcome

You are now taking part in an economics experiment. At your seat you will find an envelope—please leave this envelope closed for the time being. Inside the envelope is a slip of paper with the seat number of another participant - conversely, this participant will find your seat number in his or her envelope.

Please note: This participant is NOT the person who you came to the experiment with.

We ask you to get to know the person whose seat number is in your envelope and introduce yourselves to each other immediately afterwards. You will have two minutes for this. On the slip of paper, you will find precise instructions about where to meet this person.

Please now open the envelope and follow the instructions on the slip of paper.

Sheet 2a - Envelope (staying seated)

The person you are now going to meet has the

Seat Number [seat number X]

Please remain seated. The person with Seat Number [seat number X] will come to your seat. You will have two minutes to introduce yourselves to each other.

Sheet 2b - Envelope (walking over)

The person you are now going to meet has the

Seat Number [seat number Y]

Please take this slip with you and go to Seat [seat number Y] now. The person will be waiting for you there. You have two minutes to introduce yourselves to each other.

Sheet 3 - General Explanations

Explanations for the Experiment

This experiment is financed by various research funding institutions. By carefully reading the following explanations, you will see that, depending on your own decisions and those of other participants, you can earn a considerable amount of money. It is therefore important that you read all instructions thor-

oughly.

The instructions you have received are intended solely for your private information. **Speaking is prohibited during the experiment.** If you have any questions, please direct them to us. Failure to comply with this rule will result in exclusion from the experiment and forfeiture of all payments. If you have a question, please raise your hand. A member of the study team will then come to you and answer your question privately.

At the end of the experiment, all the amounts of money you have earned through your own decisions and those of other participants will be added up and immediately paid out to you in cash. During the experiment, we will not speak of euros but of points. Your total income will therefore first be calculated in points. The total number of points you earn during the experiment will then be converted into euros at the end, according to the exchange rate

1 point = 30 cents.

All participants will be divided into groups of 3 members each. **Apart from us, i.e., the organizers of this experiment, no one will know the group assignments.**

On the following pages, we will describe the exact procedure of the experiment.

The Decision Situation

The experiment consists of two parts. You will learn the exact procedures of these parts later. The basic structure is the same in both parts and will be explained to you in detail below. Before the actual experiment begins, there will be practice questions designed to increase your familiarity with the decision

situation.

You are a member of a **3-person group**. Each member of this group must decide how to allocate 20 points. You may either place the 20 points into a **private account**, or invest them **wholly or partially** into a project. Every point that you do not invest in the project is automatically placed in the private account.

Income from the Private Account

For every point that you place into the private account, you earn exactly one point. For example, if you place 20 points into the private account (and thus invest nothing into the project), you earn exactly 20 points from the private account. If, for example, you place 6 points into the private account, you receive an income of 6 points from the private account. *No one other than you receives any income from your private account.*

Income from the Project

All group members benefit equally from the amount you invest in the project. Conversely, you also benefit from the investments made by the other group members. The income of each member from the project is determined as follows:

$$\boxed{\text{Income from the project} = \text{Sum of contributions to the project} \times 0.5}$$

For example, if the sum of contributions by all group members to the project amounts to 40 points, then you and all other group members each receive $40 \times 0.5 = 20$ points from the project. If the three group members contribute a total of 10 points to the project, then you and all other group members each receive $10 \times 0.5 = 5$ points from the project.

Total Income

Your total income is the sum of your income from the private account and your income from the project. That is:

Income from the private account (= 20 − your contribution to the project)
Income from the project (= 0.5 × sum of all contributions to the project)
Total income

Practice Exercises

Please now answer the following practice exercises on your computer. Their sole purpose is to familiarise you with the calculation of incomes that arise from different decisions regarding the use of the 20 points. *If you have any questions, please raise your hand. A member of the study team will come to your seat and be happy to assist you.*

Sheet 4 - Phase 1

Based on the draw you completed at the very beginning, you were assigned to a 3-person group. Your two group members are therefore two randomly selected participants of this experiment, but **neither** the person with whom you came to the experiment **nor** the person you met at the start of the experiment.

The decision situation is as described at the beginning. The decisions you make will be paid out to you in cash at the end of the experiment. This part of the experiment will be conducted **only once**. As you know, you have 20 points at your disposal, which you may allocate either to your private account or to the project.

Each group member must make **two types** of contribution decisions, which we will hereafter refer to as the **unconditional contribution** and the **contribution table**.

- For the **unconditional** contribution to the project, you simply decide how many of the 20 points you would like to invest in the project. You will enter this contribution on the following computer screen:

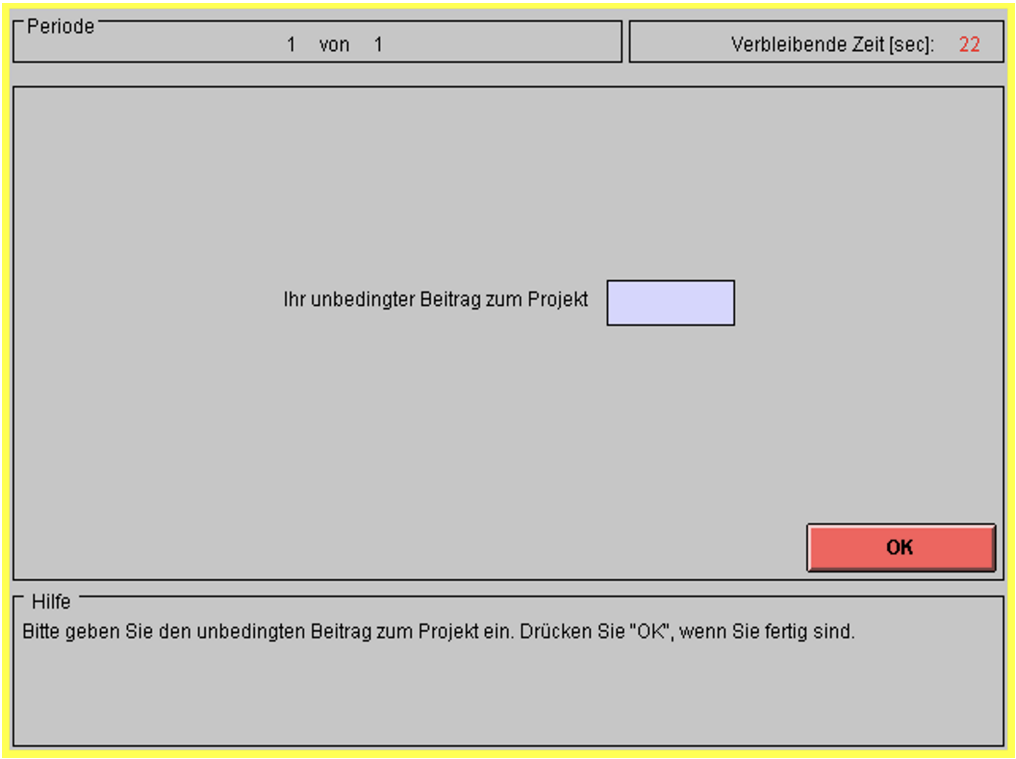


Figure A.4: Decision screen for unconditional contribution choice in Part 1. (*Figure number and caption omitted in instructions.*)

Once you have determined your unconditional contribution, press the ”OK” button.

- Your second task is to complete a **contribution table**. In this table, you must specify, **for each possible** (rounded) **average contribution of the other two group members to the project, how much you would like**

to contribute to the project. In other words, you can determine your contribution depending on how much the others contribute on average. This will become immediately clear when you view the following screen, which appears directly after you have entered your unconditional contribution.

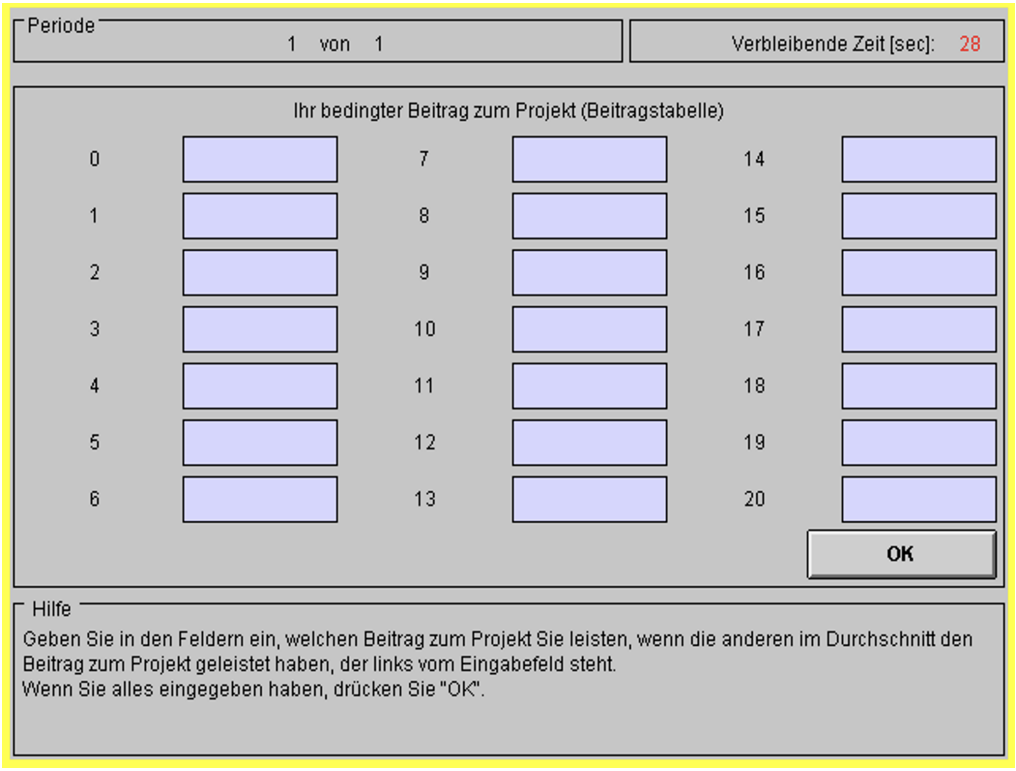


Figure A.5: Decision screen for conditional contribution choices in Part 1. (*Figure number and caption omitted in instructions.*)

The numbers to the left of the (on-screen) blue input fields represent the possible (rounded) average contributions of the **other** group members to the project. You must now simply enter into each input field how many points **you** would like to contribute to the project—under the assumption that the others, on average, make the indicated contribution. **You must make an entry in every field.** For example, you must specify how much you would contribute if the other group members on average contribute 0 points to the project; how much you would contribute if they on average contribute 1, 2, or 3 points, and so on. In each

field, you may enter any whole-number contribution from 0 to 20.

Once you have entered a value in all fields, press the ”OK” button.

After all participants in all groups have made their unconditional contribution decision and completed the contribution table, one member of each group is randomly selected. **For the randomly selected member**, only the completed **contribution table** is relevant for decisions and payments. For the **other two group members** who are not selected, only the **unconditional contribution** is relevant for decisions and payments. When you make your unconditional contribution decision and complete the contribution table, you do not yet know whether you will be randomly selected. Therefore, you must carefully consider both types of contribution decisions, since both may become relevant for you. Two examples will illustrate this.

Example 1: Suppose that **you are randomly selected, meaning that your contribution table is relevant for the decision**. For the other two members of your group, the unconditional contribution decision is relevant. Let these be 0 and 4. The average contribution of these two members is therefore 2.

If you specified in your contribution table that you would contribute 1 when the others contribute an average of 2, then the group's total contribution to the project is $0+4+1=5$. Each group member therefore earns $5\times 0.5 = 2.5$ points from the project, plus their respective income from the private account.

If instead you specified in your contribution table that you would contribute 19 when the others contribute an average of 2, then the group's total contribution to the project is $0 + 4 + 19 = 23$. Each group member therefore earns $23 \times 0.5 = 11.5$ points from the project, plus their respective income from the private account.

Example 2: Suppose that **you are not randomly selected**, so that for you and

one other member **the unconditional contribution decision is relevant**. Let your unconditional contribution be 16 and that of the other be 20. The average unconditional contribution of you and the other member is therefore 18.

If the randomly selected member specified in the contribution table that they would contribute 1 when the other two members contribute an average of 18, then the group's total contribution to the project is $16 + 20 + 1 = 37$. Each group member therefore earns $0.5 \times 37 = 18.5$ points from the project, plus their respective income from the private account.

If instead the randomly selected member specified in the contribution table that they would contribute 19 when the other two members contribute an average of 18, then the group's total contribution to the project is $16 + 20 + 19 = 55$. Each group member therefore earns $0.5 \times 55 = 27.5$ points from the project, plus their respective income from the private account.

The **random selection** of participants proceeds as follows. Each group member receives a member number between 1 and 3. The computer will then randomly draw one participant. **After** all participants have chosen their unconditional contributions and completed the contribution table, this randomly selected participant will draw a number between 1 and 3 and enter it into the computer.

If the number drawn corresponds to the number assigned to you, then your contribution table is relevant for the decision, while the unconditional contribution is relevant for the other group members. Otherwise, your unconditional contribution is relevant for you.

Sheet 5 - Phase 2

You remain assigned to randomly formed 3-person groups. Your two group members are again two randomly selected participants of this experiment, but **neither** the person with whom you came to the experiment **nor** the person you met at the beginning of the experiment.

The decision situation is once again the same as described at the beginning. The decisions you make will be paid out to you in cash at the end of the experiment. This part of the experiment will also be conducted only **once**.

Once again, you have 20 points at your disposal, which you may either place into your private account or invest in the project. Each group member must make their contribution decision. This time, however, there is no contribution table - only an unconditional contribution, which in every case is relevant for decisions and payments.

Your total income is determined in the same way as in Part 1, by summing your income from the private account and your income from the project. That is:

Income from the private account (= *20 − contribution to the project*)
Income from the project (= *0.5 × sum of all contributions to the project*)
Total income

Below, only in Treatments Anonymous, Stranger, Friend

There are two different types of 3-person groups: **observed groups** and **observer groups**.

There are an equal number of observed groups and observer groups.

You and the person you met at the beginning of the experiment are assigned to different group types. (*Stranger*)/You and the person with whom you came to the experiment are assigned to different group types. (*Friend*)

The contribution decisions are made sequentially according to group type:

- **FIRST**, each member of an **observed group** makes their contribution decision.
- **THEN**, this decision is observed by *exactly one* member of an **observer group**. Each observer sees the contribution decision of a mutually anonymous participant. **This person is NEITHER the one with whom you came to the experiment NOR the person you met at the beginning of the experiment.** (*Anonymous*)/the person **they met at the beginning of the experiment** (*Stranger*)/the person **with whom they came to the experiment** (*Friend*)

  This means that each member of an observer group will see on their screen how much a mutually anonymous participant, who is also taking part in this experiment (*Anonymous*)/ the person they met at the beginning of the experiment (*Stranger*)/the person with whom they came to the experiment (*Friend*), has contributed to the project. ONLY this ONE contribution decision will be displayed.

If you are a member of an observed group, you therefore know that your contribution to the project will be observed by a mutually anonymous person (*Anonymous*)/ the person you met at the beginning of the experiment (*Stranger*)/the person with whom you came to the experiment (*Friend*).

If, on the other hand, you are a member of an observer group, you will observe

the contribution of a mutually anonymous person who is also participating in this experiment (*Anonymous*)/ the person you met at the beginning of the experiment (*Stranger*)/the person with whom you came to the experiment (*Friend*) before you must decide on your own contribution to the project.

The following figure is intended to illustrate this once again:

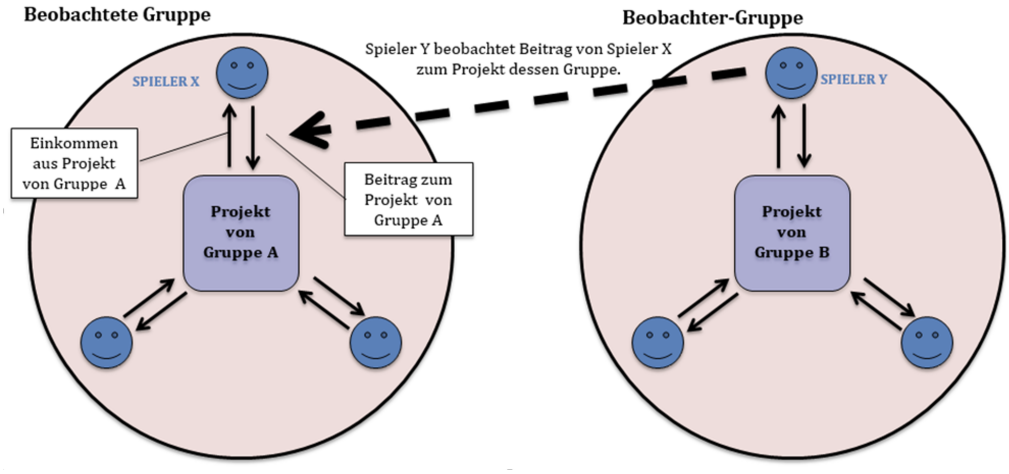


Figure A.6: Illustration of observability as shown in instructions. (*Figure number and caption omitted in instructions.*)

Note: Since observations always take place between *different* groups, only contributions are observed that do not concern the project of your own group. The observed contribution is therefore *independent* of your earnings and those of your group.

In other words, the observation - that is, the contribution of this mutually anonymous person (*Anonymous*)/ the person you met at the beginning of the experiment (*Stranger*)/the person with whom you came to the experiment (*Friend*) - plays no role in the calculation of *your* earnings.

Conversely, the contributions of the observer groups are, of course, also irrelevant for the earnings calculations of the observed groups.

Your earnings depend solely on *your* own decisions and those of *your* group members.

Whether you are a member of an observed group or an observer group was also determined by the draw you completed at the very beginning of the experiment. You will learn this in the course of the experiment.

Sheet 6 - SVO Task

In the following part, you will be randomly paired with another person participating in this experiment. The identity of this other person will remain mutually anonymous. You will make a series of decisions regarding the allocation of points between yourself and this other person. For each of the following questions, please indicate which allocation of points you prefer the most. You may select only one allocation per question.

Your decisions will determine the payment of money according to the allocation you have chosen, with

100 points = 1 Euro.

After you have chosen an allocation for all 15 questions that make up Part 3 of the experiment, one allocation will be randomly selected from your choices, and this will then be paid out. At the time you make your allocation decisions, you do not yet know which of the questions will be paid out. You should therefore consider all of your choices carefully.

There are no right or wrong answers; it is solely a matter of your personal preferences.

IOS Task

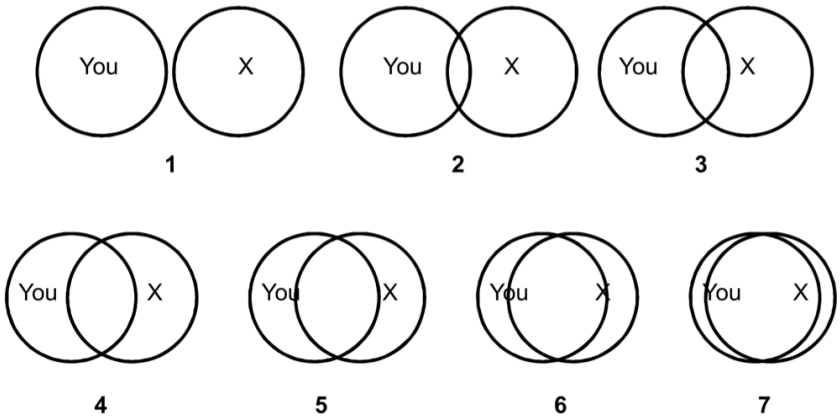


Figure A.7: IOS task. Participants were instructed to select the pair of circles which best represents how close they feel to their observation partner.

Appendix B

Chapter 2

B.1 Supplementary Tables

Table B.1: Linear probability models for strength of preferences

	Coefficient	**SE**	**t-score**	**df**	**p-value**
Intercept	0.38	0.02	19.66	53	0.000
Strength	-0.21	0.02	-13.64	206	0.000
Case 0/2	-0.15	0.03	-5.38	103	0.000
Case 1/2	-0.01	0.03	-0.26	103	0.792
Case 2/2	0.19	0.04	4.75	103	0.000
Case 0/2 × Reward	-0.01	0.03	-0.46	103	0.649
Case 0/2 × Punishment	-0.03	0.03	-0.90	97	0.370
Case 1/2 × Reward	0.03	0.04	0.76	103	0.450
Case 1/2 × Punishment	-0.03	0.04	-0.90	97	0.370
Case 2/2 × Reward	0.11	0.05	2.02	103	0.046
Case 2/2 × Punishment	0.15	0.06	2.55	97	0.012

Notes: Linear probability models for strength of preferences. Table shows coefficients of linear probability models with cluster robust standard errors in parentheses. The dependent variable takes on the value 1 if the choice is intransitive, and 0 otherwise.

B.2 Moral Problems

[Translated from German]

Swimmers

It's summer and Ralph is with his family on a private beach without a lifeguard. Suddenly he hears desperate cries from the sea. Several children have been caught in a storm and are on the verge of drowning. As Ralph is the only swimmer on site, he has to decide who to save. Those he doesn't save will drown.

According to your personal moral convictions: Who do you want Ralph to save?

1 His daughter, who had already been swept further out to sea by the storm.

2 His two nieces, who are still reasonably close to the shore.

3 Three unknown children, not far from the shore, but already with their faces under water.

HIV infection

Ken is a doctor caring for a patient suffering from HIV. This patient has told Ken in confidence that he intends to infect as many people as possible with HIV. Due to medical confidentiality, Ken sees no legal way to prevent the man from carrying out his plan.

According to your personal moral convictions: What should Ken do?

1 Do nothing due to medical confidentiality, even if the patient will infect other people with HIV.

2 Comply with medical confidentiality, but add an (undetectable) drug to the patient's medication that makes him impotent and thus prevents further infections.

3 Violate his medical confidentiality by accessing the patient's cell phone in the checkroom without being noticed and posting his intentions on the patient's social media. This would socially isolate the patient and he could not infect anyone with HIV. No one would know that Ken had posted this information.

Patients

Annabelle is a doctor. Based on a chance finding, she diagnoses a fatal illness in her patient. If he is not treated, he will die suddenly and without symptoms within the next 6 months. In his medical file, she finds a directive he signed years ago in which he expressly wishes not to be informed about this disease should it ever occur to him. At that time, the disease was incurable. However, there are now new therapies with which the disease can be cured with a 50% probability.

According to your personal moral convictions: What should Annabelle do?

1 Not inform the patient about his illness in accordance with his wishes. He would therefore die soon, but could spend the last months of his life in accordance with his wishes.

2 Inform the patient about his illness contrary to his wishes. He would then have to live with the knowledge of his illness, but could possibly be cured.

3 Breach medical confidentiality and inform a relative about the illness. This person could assess better than Annabelle whether the patient wanted to be informed about his illness or not.

Airport

Robert has a managerial role at an airport. His job is to evaluate and reorganise the random checks for dangerous objects or smuggled goods on passengers. The security staff's experience shows that passengers with certain external characteristics are more likely than others to be carrying unauthorized items.

According to your personal moral convictions: What action should Robert take?

1 Check certain groups of people more frequently. Although this is discriminatory, it increases the safety of all passengers and reduces the trade in smuggled goods.

2 Check all passengers with equal probability. This avoids discrimination, but at the expense of general security and means that smuggling of goods often goes undetected.

3 Check all passengers with equal probability. This avoids discrimination. In order to achieve the same level of general security and detection of smuggled goods as would be possible through targeted checks on these groups of people, he increases the budget for security checks. However, in order to finance this, Robert has to cancel a planned redesign of the airport to make it suitable for disabled people.

Germophobia

Roman has arranged to meet a friend for dinner one winter evening. Despite the difficult road conditions, he drives the hour or so to the restaurant. Once there, he wants to let his mother know that he has arrived safely. Unfortunately, his cell phone battery is flat. He knows that the longer she doesn't hear from him, the more worried she becomes. Asking his friend for his phone would be pointless as he suffers from a pronounced form of germophobia and is terrified of germs on his phone.

According to your personal moral convictions: What should Roman do?

1 Roman ignores his friend's phobia and waits until he goes to the toilet. Then he takes his friend's cell phone, informs his worried mother, deletes the call and puts it back in the same place. His friend will never know that he used it.

2 Roman leaves his mother in the dark.

3 Roman steals a bottle of disinfectant spray from an unattended handbag in the checkroom. This allows him to use his friend's phone to reassure his mother and then simply disinfect it.

Covid-19

The country where Ben lives is so badly affected by the coronavirus that the intensive care units are overloaded and there are not enough life-saving ventilators available. It is therefore necessary to decide which patients should be treated first. Ben is leading an interdisciplinary team that has been tasked with drawing up triage guidelines for the clinics. These are intended to specify which patients are to be treated and with what priority.

According to your personal moral convictions: Who should Ben's team give higher priority to?

1 Younger people.

2 People who require intensive care for a shorter period of time.

3 People who have been infected through no fault of their own (not through negligence or disregard of the corona regulations).

Grandparents

Christian is a virologist. During a pandemic, the number of infected people is rising rapidly and a lockdown is imminent. Christian's sister is afraid of being locked up in a small city apartment with her family for weeks on end due to the impending lockdown. That's why she plans to make a quick trip with her children to her parents, who live in a large house with a garden. Christian's parents are very much looking forward to having the family with them for several weeks. At the same time, children are possible carriers and older people are at high risk. Christian is a virologist and knows that his sister's plan threatens his parents' health.

According to your personal moral convictions: What should Christian do?

1 Accept the wishes of his sister and parents and let the visit happen.

2 Clearly tell his sister and parents that he does not agree with this. This would result in a major family dispute and would be a great burden for his sister and parents. However, his sister would cancel the visit.

3 Falsely claim that there is an as yet unpublished study showing that almost all children are now asymptomatic carriers. Consequently, the grandparents would very likely be infected if they came into contact with his sister's children. His sister and parents would therefore voluntarily refrain from visiting.

Emergency

Hermione is a doctor in an emergency room. A patient comes to her with a bleeding wound. He belongs to a religious community that forbids blood transfusions. If a vital blood transfusion is necessary, he refuses it. It soon becomes clear that the wound does indeed require a blood transfusion. The patient is now unconscious.

According to your personal moral convictions: What should Hermione do?

1 Respect the patient's wishes and let him die.

2 Administer the blood transfusion against the patient's wishes, save his life and inform him truthfully about the treatment. He would go on living in the knowledge that he had received a blood transfusion contrary to his religious beliefs.

3 Administer the blood transfusion against the patient's wishes, save his life, but do not inform him later about the treatment. He would continue to live in the belief that he had not received a blood transfusion.

Lung transplant

Emil is a doctor in a lung transplant clinic where patients are currently waiting urgently for a donor lung. As there is currently only one donor lung available, Emil has to decide who gets the organ and who will inevitably die.

According to your personal moral convictions: Which patient should Emil give the donor lung to?

1 A 20-year-old smoker who is not an organ donor himself.

2 A 45-year-old non-smoker who is not an organ donor himself.

3 A 70-year-old non-smoker who is an organ donor himself.

Abortion

Laura suffers from severe alcohol addiction. Laura recently found out that she is pregnant. Her partner, the father of the child, has already clearly signalled to her that he does not want a child because of Laura's alcohol dependency. Laura's ongoing alcohol consumption means that damage to the child cannot be ruled out. It would still be early enough for a legal abortion. But she wants to have the child herself.

According to your personal moral convictions: What should Laura do?

1 Have and raise a child.

2 Have an abortion.

3 Have the child and give it up for adoption.

Affair

Peter finds out by chance that his best friend is having an extramarital affair. The friend lives with his wife and their two young children.

According to your personal moral convictions: What should Peter do?

1 Tell the wife about the affair and cause the marriage to break up.

2 Force his friend to end the affair immediately by threatening to expose it. The affair would continue to be concealed from the wife and the marriage would continue.

3 Stay out of the affair. He would passively cover up the affair and conceal it from his wife.

Wedding cake

Ronja is an employee in a patisserie. A vegan bride and groom come to collect their wedding cake. Ronja notices that the cake was also made with animal products. The new employee had made a mistake when taking the order.

According to your personal moral convictions: What should Ronja do?

1 Honestly admit the mistake to the wedding couple. The wedding would have to take place without a cake. After a complaint from the wedding couple, the owner of the cake shop would fire the new employee.

2 Hide the mistake from the wedding couple. The bride and groom and the wedding guests would eat the cake and no one would notice that it was not vegan. The new employee could keep his job.

3 Falsely claim to the wedding couple that the fridge was out of order due to a power cut, which is why the cake was now inedible. The wedding would therefore have to take place without the cake. The new employee's mistake would remain undetected and he could keep his job.

B.3 Classical Moral Problems

[Translated from German]

Train

Anne happens to notice that five track workers are about to be hit and killed by an approaching, out-of-control train. The track workers are all wearing hearing protection and will only notice the train when it is already too late. As the track workers are some distance away from Anne, she has no way of warning them. The only way to save the five track workers would be to move a switch and divert the train to another track. However, there is another track worker on this track who also cannot be warned and would therefore die if Anne moved the points.

According to your personal moral convictions: What should Anne do?

1 Change the switch.

2 Do not change the switch.

Bridge

Hanno is standing on a railroad bridge for pedestrians. An approaching train gets out of control and races towards five track workers. As the track workers are wearing hearing protection, they are unaware of the train and cannot be warned. Hanno sees a tall, corpulent man on the bridge railing. Hanno could give the man a push so that he would fall onto the tracks between the approaching train and the five track workers, which would stop the train in time.

This would be the only way to save the five track workers, but the man would die in the process.

According to your personal moral convictions: What should Hanno do?

1 Push the man off the bridge and rescue the five track workers.

2 Don't push the man off the bridge. The five track workers would then die.

Organ donation

Hugo is a surgeon. Five patients are currently waiting in his clinic for a vital donor organ. Each of the five patients needs a different organ. As there is no organ donor in sight, the five patients are all about to die. A healthy man has registered for a check-up this afternoon. Hugo could kill the man unnoticed, remove his organs and use them to save the lives of the five patients.

According to your personal moral convictions: What should Hugo do?

1 Kill the healthy man, remove his organs and save the five patients.

2 Do nothing and let the five patients die.

B.4 Instructions & Payout Info

Instruction screens are shown in order of appearance. Each screen was shown in all treatments, unless indicated otherwise. Treatment differences are highlighted in red.

[Translated from German]

Screen 1

General Procedure of the Experiment

This experiment consists of **one (*No Info*)/two (*Only Info*)/three (*Reward, Punishment*) stage(s)**, in which you will make several decisions in each stage.

(For each stage,) you have a predetermined amount of time. The remaining time is displayed to you in the upper right corner. The time is set so that it should be sufficient to process all decisions. As soon as you have completed all your decisions in a stage before the end of the time limit, the experiment immediately continues. If, at the end of the time, some decisions remain unanswered, they will be randomly generated by the computer and the experiment automatically continues.

For each of your decisions, you will receive an amount of 0.35 Experimental Currency Units (ECU). In addition, you will receive a flat fee of 5.00 Euros for your participation. Your total income is the sum of the ECU from all parts converted into Euros (1 ECU = 1 Euro) and the flat participation fee.

In this experiment, we will share some of your decisions with other participants, but you will remain anonymous. Similarly, you will be shown some of the decisions of other participants. If other participants have not made a decision

within the given time, you will be informed that these were randomly generated decisions.

Screen 2

Overview

According to your personal moral conviction: Which option will you choose?

On your screen, you will successively see descriptions of moral dilemmas, each with two possible courses of action. Please decide for each dilemma which of the two options most closely corresponds to your personal moral conviction.

You will be randomly assigned to groups of 3 people. The other members of your group will also make such decisions.

Treatments *Reward* and *Punishment* only: Later, for each dilemma, a person outside your group will be shown which options you and your group members have chosen. Based on these decisions, this person must - separately for each dilemma - determine one member of your group to receive a reward (*Reward*)/deduction (*Punishment*) of 10 ECU.

At the end of the experiment, one of these moral dilemmas will be randomly selected for your group. The group member who was assigned the reward (*Reward*)/deduction (*Punishment*) of 10 ECU in this dilemma, based on the chosen option as judged by the evaluator, will have 10 ECU added (*Reward*)/deducted (*Punishment*).

In the following, it (*No Info*)/the two stages (*Only Info*)/the three stages (*Reward, Punishment*) will be described in more detail.

Screen 3a

After the description of a moral dilemma, you will see two possible courses of action, A and B. You decide which of the two options most closely corresponds to your personal moral conviction. To do so, click on the corresponding answer.

After each decision, we ask you to indicate how strongly you prefer the option you have chosen. After your decision, a bar will appear below the options. Please try this by clicking on one of the options (A or B). If you then click on the bar that appears, you can use the diamond symbol to indicate how certain you are about your decision.

You will make such decisions one after another for 27 moral dilemmas.

According to your personal moral conviction: Which option will you choose?

[clickable buttons:] *A* or *B*

[slider:] *certainly A - indifferent - certainly B*

Screen 3b - Treatments *Only Info*, *Reward*, *Punishment*

In Stage 2, you will again see descriptions of moral dilemmas. Please indicate once more which of the two courses of action most closely corresponds to your personal moral conviction, and how certain you are about your choice.

This time, before each decision, you will also see how the other two members of your group decided in this dilemma. The others have already gone through this dilemma in Stage 1. At that time, they had no information about the decisions of others. The options chosen by your group members will be shown to you for each dilemma, again in random order.

An example of this presentation is shown below. Please also try out this stage by clicking on the fields A or B in the lower row. Your chosen option will then appear in the gray field of the upper row.

You will go through this decision situation for 12 moral dilemmas.

These are the decisions of your group members:

chosen options of group members

According to your personal moral conviction: Which option will you choose?

[clickable buttons:] *A* or *B*

[slider:] *certainly A - indifferent - certainly B*

Screen 3c - Treatments *Reward*, *Punishment*

In Stage 3, the decisions that you and your group members made in a dilemma are transmitted to a member of another group for evaluation. This evaluator must, based on the respective decisions, mark one person in your group to receive a bonus (*Reward*)/deduction (*Punishment*) of 10 ECU. The evaluator does not know which decisions were made in Stage 1 and which in Stage 2. In addition, the positions in which the chosen options of the 3 group members appear are randomly determined again for each dilemma. You too will evaluate the decisions of another group.

At the end of the first part, one dilemma will be randomly selected for each group. The person marked in this dilemma will then receive the bonus (*Reward*)/deduction (*Punishment*). Below, you can see the evaluation situation that results from the example in Stage 2. One member of your group chose Option A in Stage 1 and another chose Option B. Your decision from Stage 2 appears

on the right. Now take on the role of the evaluator and click on one of the three chosen options.

In this third part, you will make a total of 12 evaluations.

Who shall receive the bonus (*Reward*)/deduction (*Punishment*)?

[clickable buttons:] *A* or *B* or *B*

Screen 4

Control Questions

Please check, using the control questions, whether you have understood the instructions correctly. Your answers have no effect on the experiment.

1) How many stages are there in this experiment?

○ 1 ○ 2 ○ 3

2) The decisions of how many other participants do you see in Stage 2?

○ 2 ○ 5 ○ 9

3) In which stage did the other participants go through this decision situation?

○ 1 ○ 2 ○ 3

4) What happens to the unanswered decisions of a person if the time for a stage has expired? They are ...

○ made up later ○ randomly generated

5) What happens if the displayed decision of another participant in Stage 2 was generated by the computer?

○ I am informed ○ I am not informed

[clickable button:] Show solutions

Screen 5

Conclusion

At the end, you will receive an overview of your income from the experiment. The money will be transferred to the account you provided.

Working on the moral dilemmas may occasionally cause emotional stress. Your participation is voluntary, and you may withdraw at any time.

Your payment is calculated from all the decisions you have made. If you wish to withdraw early, we kindly ask you to keep your browser open so that, at the end of the experiment (when the other participants have made their decisions), you can enter your bank account details for your payment.

If you have questions, you can now contact us via **BigBlueButton**.

If you agree to participate in this study, please click the ”START” button.

Payout Info (End of Experiment)

Payout Overview

You receive a flat rate of 8.50 ECU.
In Stage 1 you processed 4 decisions and therefore receive $4 \times 0.35 = 1.40$ ECU.
In Stage 2 you processed 7 decisions and therefore receive $7 \times 0.35 = 2.45$ ECU.
In Stage 3 you processed 13 decisions and were selected for the deduction. For that you receive $13 \times 0.35 - 10 = -5.45$ ECU.
In total, you therefore receive 6.90 ECU for your participation.

A short questionnaire will now follow. At the end of the questionnaire you will be asked to enter your IBAN. We will transfer your payout to the specified account within the next few days.

B.5 Decision Screens

Without social influence

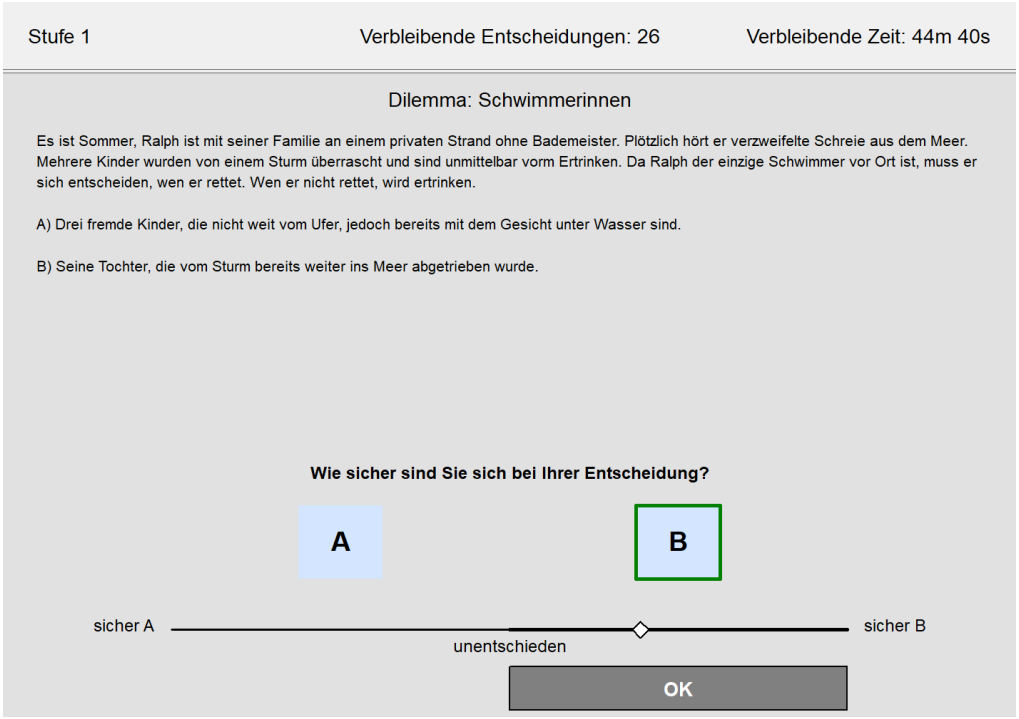


Figure B.1: Selection screen for decisions without social influence. (*Figure number and caption omitted in instructions.*)

English Translation:

It's summer and Ralph is with his family on a private beach without a lifeguard. Suddenly he hears desperate cries from the sea. Several children have been caught in a storm and are on the verge of drowning. As Ralph is the only swimmer on site, he has to decide who to save. Those he doesn't save will drown.

A Three unknown children, not far from the shore, but already with their faces under water.

B His daughter, who had already been swept further out to sea by the storm.

According to your personal moral conviction: Which option will you choose?

[clickable buttons:] *A* or *B*

[slider:] *certainly A - indifferent - certainly B*

With social influence

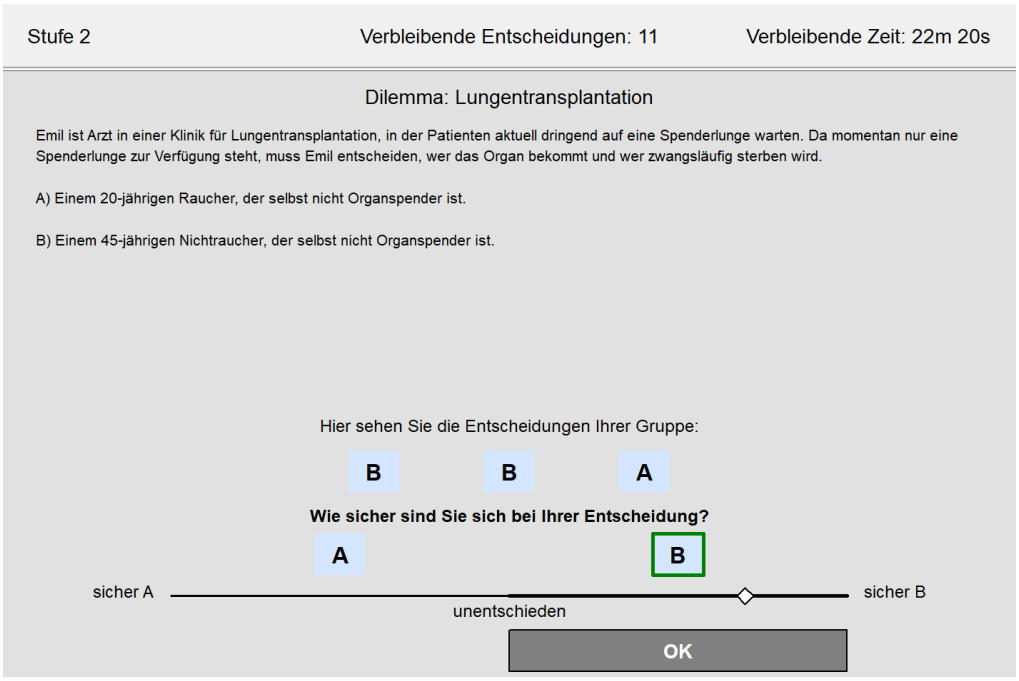


Figure B.2: Selection screen for decisions with social influence. (*Figure number and caption omitted in instructions.*)

English Translation:

Emil is a doctor in a lung transplant clinic where patients are currently waiting urgently for a donor lung. As there is currently only one donor lung available, Emil has to decide who gets the organ and who will inevitably die.

A A 20-year-old smoker who is not an organ donor himself.

B A 45-year-old non-smoker who is not an organ donor himself.

These are the decisions of your group members:

chosen options of group members

According to your personal moral conviction: Which option will you choose?

[clickable buttons:] *A* or *B*

[slider:] *certainly A - indifferent - certainly B*

Evaluation in Treatment *Reward*

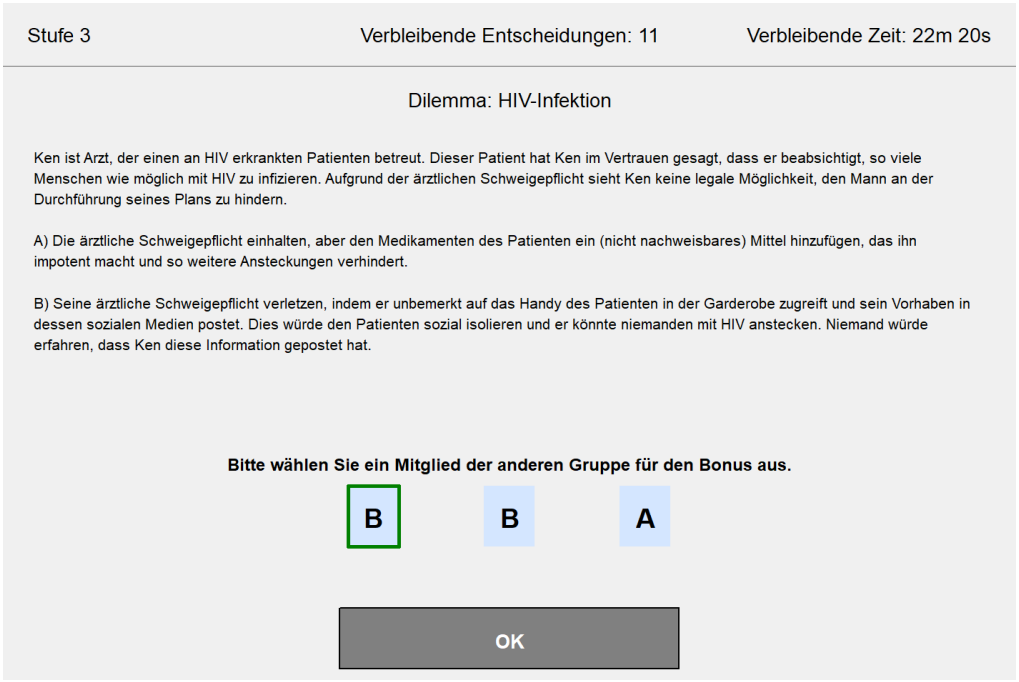


Figure B.3: Selection screen for reward decisions. (*Figure number and caption omitted in instructions.*)

English Translation:

Ken is a doctor caring for a patient suffering from HIV. This patient has told Ken in confidence that he intends to infect as many people as possible with HIV. Due to medical confidentiality, Ken sees no legal way to prevent the man from carrying out his plan.

A Comply with medical confidentiality, but add an (undetectable) drug to the

patient's medication that makes him impotent and thus prevents further infections.

B Violate his medical confidentiality by accessing the patient's cell phone in the checkroom without being noticed and posting his intentions on the patient's social media. This would socially isolate the patient and he could not infect anyone with HIV. No one would know that Ken had posted this information.

Who shall receive the reward?

[clickable buttons:] *A* or *B* or *B*

Evaluation in Treatment *Punishment*

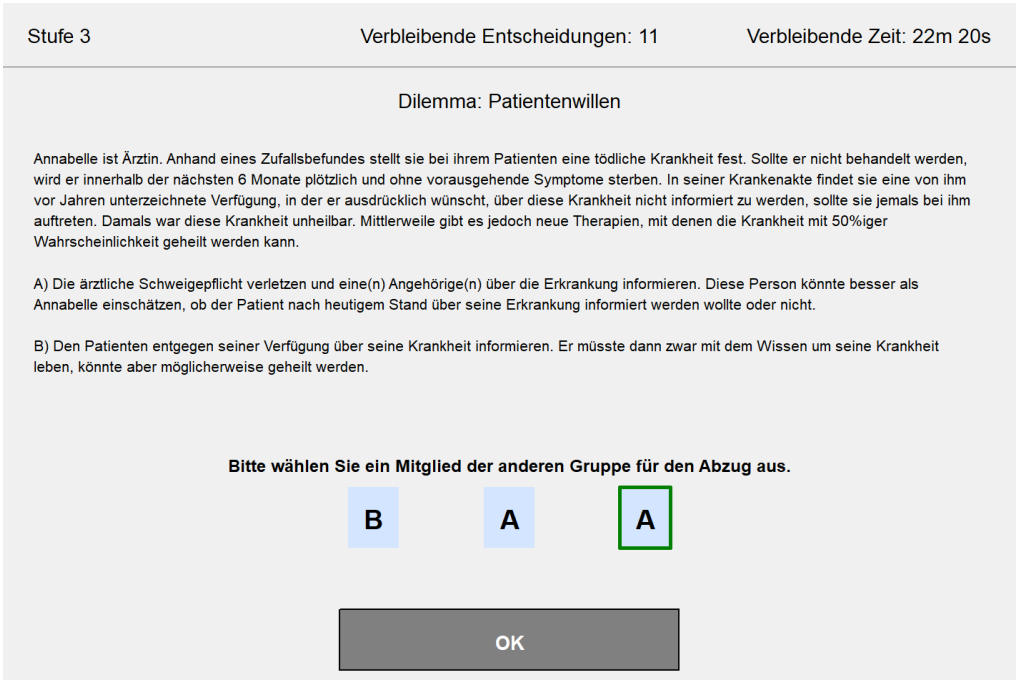


Figure B.4: Selection screen for punishment decisions. (*Figure number and caption omitted in instructions.*)

English Translation:

Annabelle is a doctor. Based on a chance finding, she diagnoses a fatal illness in her patient. If he is not treated, he will die suddenly and without symptoms within the next 6 months. In his medical file, she finds a directive he signed years ago in which he expressly wishes not to be informed about this disease should it ever occur to him. At that time, the disease was incurable. However, there are now new therapies with which the disease can be cured with a 50% probability.

According to your personal moral convictions: What should Annabelle do?

A Breach medical confidentiality and inform a relative about the illness. This person could assess better than Annabelle whether the patient wanted to be informed about their illness or not.

B Informing the patient about their illness contrary to their wishes. He would then have to live with the knowledge of his illness, but could possibly be cured.

Who shall receive the deduction?

[clickable buttons:] *A* or *B* or *B*