Université Paris 1 Panthéon-Sorbonne
École doctorale d'Économie (ED 465)
Centre d'Économie de la Sorbonne (CES)

18/12/2025

Ph.D. Thesis

Submitted for the Degree of Doctor of Philosophy in Economics

Visibility and Social Closeness: How Information Shapes Beliefs, Coordination, and Third-Party Responses

Presented by

Irving Argaez Corona

Under the supervision of

Jean-Christophe Vergnaud & **Béatrice Boulu-Reshef**

Jury

Agnès Festré, Université Côte d'Azur	reviewer
Fabio Galeotti, Université de Lyon	reviewer
Camille Chaserant, Université Paris 1	referee
Lusine Grigoryan, University of York	referee

Abstract

Socialisation is a fundamental aspect of human behaviour. Through interactions, individuals develop a sense of affiliation and understand their place in social structures. The outcomes of these interactions, however, depend a great deal on what is known about others and on how such information is internalised in decision-making. This thesis examines how making social information visible reshapes beliefs and behaviour in strategic settings. Its purpose is twofold: (1) to test whether revealing social information impacts expectations, behaviour, and third-party enforcement, and (2) to trace whether these effects operate through strengthened preferences for similar others.

Drawing on behavioural economics, social psychology and experimental game theory, the project articulates a cohesive framework for how social information operates along a single pipeline-from making information salient, to shaping beliefs and primary behaviour, to third-party enforcement. Using incentivised laboratory and online experiments, the thesis investigates whether social closeness-modelled via personality labels, and objective and subjective measures of socioeconomic status and political affiliation-shapes strategic decisions and preferences for socially close counterparts, and whether this, in turn, alters coordination and third-party enforcement.

The thesis is structured in three chapters. **Chapter 1** analyses how making personality information visible shapes beliefs and strategic coordination. In a repeated stag-hunt with belief elicitation, information disclosure improves strategic coordination and initially tilts expectations in favour of more trusting counterparts, largely driven by expectations about others' behaviours and learning from repetition. **Chapter 2** investigates whether objective or subjective social closeness in socioeconomic status or political affiliation increases dishonest reporting when payoffs are independent and monitoring is passive. In a repeated Die-under-the-cup task, findings show that closeness reliably shifts expectations, yet does not robustly increase misreporting. **Chapter 3** examines whether observers punish and reward differently across social distance and status once suspicion signals are available. The chapter shows that while increased cheating is associated with increased punishment and lower rewards, these decisions are selectively biased, as leniency and severity towards similar others vary with social distance and with socioeconomic status.

Keywords: social information; disclosure; social closeness; social distance; socioeconomic status; political preferences; norm enforcement; observability; beliefs; coordination; economic decisions; suspicions of cheating; Die-under-the-cup task; stag-hunt game; in-group favouritism; expectations; personality traits; agreeableness; social preferences; decision-making

Résumé

La socialisation est un aspect fondamental du comportement humain. À travers leurs interactions, les individus développent un sentiment d'appartenance et situent leur place au sein des structures sociales. Cependant, l'impact de ces interactions dépende largement de ce que l'on sait des autres et de la manière dont ces informations sont internalisées dans la prise de décision. Cette thèse examine comment le fait de rendre visibles des informations sociales reconfigure les croyances et les comportements dans des contextes stratégiques. Elle poursuit un double objectif : (1) tester si la divulgation d'informations sociales influence les attentes, les comportements et l'application des normes par des tiers, et (2) retracer si ces effets opèrent à travers un renforcement des préférences pour les individus similaires.

S'appuyant sur l'économie comportementale, la psychologie sociale et la théorie des jeux expérimentale, le projet propose un cadre cohérent décrivant le fonctionnement de l'information sociale le long d'un même processus — du dévoilement de l'information, à la formation des croyances et des comportements primaires, jusqu'à l'application des normes par des tiers. À l'aide d'expériences incitatives en laboratoire et en ligne, la thèse étudie si la proximité sociale—modélisée à partir d'étiquettes de personnalité et de mesures objectives et subjectives de niveau socioéconomique et d'affiliation politique—façonne les décisions stratégiques et les préférences pour des partenaires socialement proches, et si cela, en retour, modifie la coordination et l'application des normes par des tiers.

La thèse est structurée en trois chapitres. **Le chapitre 1** analyse comment la visibilité d'informations sur la personnalité façonne les croyances et la coordination stratégique. Dans un jeu type *stag-hunt* répété, le dévoilement d'information améliore la coordination stratégique et oriente initialement les attentes en faveur de partenaires plus confiants, un effet largement porté par les anticipations concernant le comportement d'autrui et par l'apprentissage au fil des répétitions. **Le chapitre 2** examine si la proximité sociale objective ou subjective en termes de niveau socio-économique ou d'affiliation politique augmente la malhonnêteté dans les déclarations lorsque les gains sont indépendants et la surveillance passive. Dans une tâche répétée du type « Die-under-the-cup », la proximité modifie de manière fiable les attentes, mais n'augmente pas de manière significative le faux reporting des gains. **Le chapitre 3** examine si les observateurs punissent et récompensent différemment en fonction de la distance sociale et le statut lorsque des signaux de suspicion sont disponibles. Le chapitre montre que, bien qu'ne hausse de la tricherie est associée à une augmentation des punitions et à une diminution des récompenses, ces décisions demeurent sélectivement biaisées : la clémence et la sévérité envers des individus similaires varient selon la distance sociale et selon le statut socio-économique.

Keywords: information sociale ; divulgation ; proximité sociale ; distance sociale ; niveau socio-économique ; préférences politiques ; application des normes ; observabilité ; croyances ; coordination ; décisions économiques ; soupçons de tricherie ; tâche du "stag-hunt" ; favoritisme endogroupe ; attentes ; traits de personnalité ; amabilité ; préférences sociales ; prise de décision

Acknowledgements

I find it ironic that an often solitary experience was actually shared by an entire network of people who supported me along the way.

My gratitude goes to my thesis supervisors, Jean-Christophe Vergnaud and Béatrice Boulu-Reshef, for guiding me along this process and for their invaluable feedback for this project. Thank you for your patience during what often felt like a never-ending learning curve, and for taking me under your wing despite not having an academic background. Your support was essential to the completion of this endeavour.

I also thank Agnès Festré, Camille Chaserant, Fabio Galeotti and Lusine Grigoryan for agreeing to be part of my thesis jury and for the time they dedicated to reading and discussing this work. Agnès and Fabio: your feedback from the pre-defence was crucial in making this project better, thank you for that.

To the people at *Maison des Sciences Economiques* for their support during these years: Vincent de Gardelle and Marie-Pierre Fayant for participating in my thesis committee; Alexiane Chapron, Amelie Collin, Eric Delogu, Frédéric Busson and Rachel Creppy for your friendliness and warmth during my time at CESdoc, and especially to Emilie Roux for this opportunity. A big thank you to Maxim Frolov, your support was essential in making the experiments happen, thank you for making sure everything ran smoothly.

I thank the National Council of Science and Technology of Mexico (CONACYT) for funding my thesis project. I was fortunate to receive publicly-funded support for my postgraduate studies on two occasions—CONACYT's scholarship programmes allowed many students from non-privileged backgrounds to pursue their studies, and it is regrettable that the current administrations have decided to deprive future generations of this opportunity.

I also thank Université Paris 1 for providing me with the opportunity to venture into teaching. This was a new, enticing experience that provided me with great learning opportunities-and with the chance to say that I taught at THE Sorbonne. Thanks to the professors who trusted me with their courses: Arnold Njiké, Christophe Ramaux, Corinne Perraudin, Liza Charroin, Mathieu Leduc and Sandra Poncet.

I thank Pilar Ostos Centina and Claudia Rivera Fuentes, two of my undergraduate professors, for their support, mentorship and for helping me with my PhD applications.

Thanks to my peers for always making me feel welcome, for your kindness and for sharing this experience with me: Bin, Clémentine, Hélène, Jean-Michel, Justine, Laurence, Lily, Nina, Olivier, Quentin, Sharon.

Pursuing a PhD was a personal goal of mine for a long time, though I often wondered if it was truly the right path. Leaving behind stability, my country, and my career was not easy. Yet, along this path, I met a remarkable group of people who made me realise I was exactly where I needed to be. I want to thank *mis mexicanos* for their love and support: Adri, Lou, Marianne, Marie, Raph, Simon, Thom, Vic-you all embody the best of both countries. To the *famille* Labeyrie: Anna, Jean-Marc and Julie, for their warmth and for giving me the kind of welcome anyone could dream of. Thank you Jade, for being a great flatmate and an even better listener. To Aurélien and Olivier, for your friendship and all the experiences we've shared over these years. And to Caro, for being a piece of home in France.

To Andrea and Dan, my siblings from another family—we go back so many years, and none of us could have imagined that life would bring us to the same country after all this time. Thank you for always listening, for the laughs and the jokes, and for your words of comfort. There were moments when I felt like throwing myself into the Seine, but you were always there to pull me back.

To my family—my greatest privilege in life is having the family I have. Mum, Dad, Khim, thank you for your unwavering support, your love, and for always reminding me that I can achieve anything I set my mind to. During this time, your texts, your phone calls, your visits, and my trips back home, filled me with positivity and gave me the motivation to keep moving forward.

And thank you Matt. Your support, your beautiful energy, your positivity, your smile, your love —you make life so much better. Thank you for being my home and my life companion, I love you to Japan, the Moon, Pluto and back.

Contents

General Introduction

Social interactions rely on information about others to guide trust, coordination and norm enforcement. When such information becomes visible, it can shape how people form beliefs, make strategic decisions, and enforce social norms. Understanding these processes is central to behavioural economics and social psychology, as it reveals how perceptions of similarity and identity can promote cooperation but also generate bias.

This thesis examines how making social information visible reshapes beliefs and behaviours in strategic contexts, and whether these effects depend on preferences for similar others. Across three experimental studies, it investigates how identity-relevant cues, such as personality traits, socioeconomic status, and political affiliation, shape behaviour and belief formation in strategic settings.

Motivation and research questions

Learning about other people's identity is crucial in strategic interactions. The latest wave of the *World Values Survey (WVS)* reports 75% of respondents in its *Social capital, trust and organisational membership* section consider being "very careful in dealing with people" a necessary trait. When disaggregating trust by social groups (e.g. neighbours, acquaintances, other nationalities), none of these reached the 25% benchmark, with the notable exception of families (77%) (see Haerpfer et al. (2020)). Similarly, the OECD's 2024 *Survey on Drivers of Trust in Public Institutions* reports an upward trend in the share of people with low or no trust in their national government (44%), driven by factors such as socioeconomic conditions, gender or demographic characteristics (see OECD (2024)).

This data illustrates the complex reality facing trust in interactions, which often pervade social life. Evidence suggests that, in order to obtain cooperative outcomes, it is necessary to trust others and to believe that others trust us in return (Hilbig et al., 2013; Rothstein and Eek, 2009). While most people display preferences for trustworthy individuals, navigating everyday interactions paints a more complex reality of how trust and cooperation dynamics work (Barranti et al., 2016; Gächter and Schulz, 2016; Rothstein, 2011). Addressing how these dynamics come to be, as well as the mechanisms facilitating them, is a persistent focus for behavioural and experimental economics.

In this sense, similarity cues such as shared personalities, socioeconomic background or political affiliation, help individuals anticipate others' behaviour. Evidence suggests that when information about others is scarce, people rely on assortative expectations, assuming that others think and act similarly to themselves (Beer and Watson, 2008; Thielmann et al., 2020). These assumptions serve as cognitive shortcuts that enhance perceived predictability and foster a sense of mutual understanding in interactions (Martin, 2015). However, the same cues can also trigger bias, in-group favouritism, and selective enforcement of norms, creating a fundamental tension whereby information that facilitates coordination may simultaneously undermine fairness. When information about others is disclosed, beliefs may become motivated and asymmetric (Baccini and Hartmann, 2022), cooperation with perceived similar partners encouraged (Chierchia and Coricelli, 2015; Lönnqvist et al., 2021; Rubinstein and Salant, 2016) and perceptions of others shaped by self-related beliefs (Cooper and Withey, 2009).

Against this backdrop, the overarching research question in this thesis is twofold:

1. How does the visibility of social information impact beliefs, behaviour and third-party responses in strategic interactions?

2. Do these effects depend on preferences for similar others?

The thesis addresses these questions by developing a unified framework that integrates the salience of social information, cues of social closeness, and selective responses to others' behaviour. Across three experimental chapters, it investigates how visible information about others—such as personality traits, socioeconomic background, or political affiliation—shapes coordination, honesty, and third-party reactions to (non)compliance. In sum, these studies aim to shed light on the dual role of social information: as a mechanism that facilitates cooperation and mutual understanding, yet also one that may foster bias and leniency toward similar others.

Theoretical mechanisms

This section outlines the theoretical mechanisms underpinning this framework and describe how each chapter contributes to addressing the broader research questions. Each experiment manipulates the visibility of social information and measures how this affects belief formation, behaviour, and third-party responses.

1.1 Disclosing information in strategic interactions

Information plays a central role in strategic interactions. What individuals know—and know that others know—helps align expectations and reduce uncertainty, facilitating coordination toward mutually beneficial outcomes. Making relevant information common knowledge can thus speed convergence to efficient equilibria, particularly in repeated settings where disclosure supports learning and trust formation (De Freitas et al., 2019; Fehr et al., 2008; Thomas et al., 2014; Van Huyck et al., 2018).

However, information disclosure can also act as a social signal. Visible traits and labels activate social categorisation and in-group bias, shaping perceptions of similarity and difference (Akerlof and Kranton, 2000; Tajfel and Turner, 1979). As a result, transparency may simultaneously enhance predictability and foster discrimination. For example, revealing low prosociality can reduce cooperation overall, while making group identities common knowledge can trigger preferential or exclusionary behaviours (Balliet et al., 2014; Bernard et al., 2016; Drouvelis and Georgantzis, 2019; Ruzzier and Woo, 2023).

These two perspectives: information as a coordination device and information as a social signal, are often examined separately. In practice, however, they coexist and can interact. The same information that improves coordination, can also activate biases that undermine fairness. Understanding how these dynamics intertwine is particularly relevant in trust-sensitive environments, where expectations are fragile and identity cues can quickly shape behaviour.

This thesis addresses this duality experimentally. We systematically manipulate what participants know about themselves and their counterparts to analyse how information visibility

shapes beliefs, expectations, and strategic decisions. In **Chapter 1**, the disclosed information concerns trait labels categorising each participant as either "trusting" or "mistrusting." Visibility is varied across four conditions: *No information*: neither participant knows any type label; *Private information*: participants only know their own type; *Public information*: participants only know their counterpart's type; *Full information*: both participants know both labels.

This design allows us to test whether information disclosure facilitates coordination by reducing uncertainty, or whether it instead triggers social categorisation and selective cooperation. **Chapter 2** and **Chapter 3** extend this logic to socioeconomic and political information, testing how identity-relevant cues influence cheating when reporting outcomes and third-party responses to suspicions of cheating.

1.2 Social closeness and preferences for similar others

Developing a sense of belonging is an inherent human behaviour. Sharing traits, common identities or perceived affiliation increase the perceived relatedness that bind people together. Traditionally, economic theory has related strong feelings of closeness to higher trust and coordination, both in naturally occurring ties and in laboratory settings that induce minimal identities (Bicchieri et al., 2022; Jansson, 2015; Régner and Monteil, 2007; Rubinstein and Salant, 2016). However, these same cues can shift standards across group boundaries: in-group favouritism and identity salience may foster selective leniency toward close others and stricter scrutiny of distant others, especially in competitive or resource-scarce environments (Bilancini et al., 2020; Chae et al., 2022; Leidner et al., 2010).

One manifestation of these boundary effects is dishonesty. Individuals are more likely to condone or enact rule-bending that benefits close others, and to evaluate such behaviour more positively (Conrads et al., 2013; Gross et al., 2018; Hoffmann, 2013; Jordan et al., 2024; Korbel, 2016; Leidner et al., 2010; Weisel and Shalvi, 2015; **?**). Laboratory and field evidence shows that people extend greater leniency to close others and apply stricter scrutiny to distant ones for comparable signals (Anvari et al., 2019; Ellemers et al., 2008; Rullo et al., 2024; Waytz et al., 2013).

This thesis treats dishonesty as an illustrative case within a broader pattern in which social closeness shapes expectations and reshapes attitudes towards observed cheating. In the remainder of the thesis we examine these mechanisms with objective and subjective conceptualisations of closeness. Chapter 2 tests whether social closeness (socioeconomic status and political affiliation) shifts expectations and own reporting behaviour under neutral incentives, independent payoffs and passive observation. This isolates informational and identity channels from reciprocity or peer pressure. Chapter 3 focuses on how socioeconomic closeness shapes third-party punishment and reward, decomposing third-party decisions while holding statistical signals of suspicion of dishonesty constant.

1.3 Modelling social closeness

Social Identity Theory posits that an individual's identity is derived from their awareness of group membership and the emotional value they attribute to that affiliation (Tajfel and Turner,

1979). The literature underpins that traits and characteristics contribute to identity formation to different degrees.

While visible characteristics (e.g., gender, nationality, racial background) have long dominated experimental work, recent research emphasises subjective cues, such as beliefs, perceived status and personality traits, as additional bases for relatedness and belonging. In this thesis we model social closeness using two mechanisms: socioeconomic status (SES) and political affiliation, each captured in an objective and subjective form.

For SES, the objective measure corresponds to the average income in a participant's locality of residence, whereas the subjective measure reflects participants' self-reported income relative to the departmental average. For political affiliation, the objective measure corresponds to the political affiliation of the elected deputy in their department, while the subjective measure captures self-reported preferences across major political parties. Objective indicators thus anchor identity cues in contextual and plausibly exogenous environments, while subjective indicators capture self-perceived identification and personal alignment.

This modelling choice reflects the growing recognition of place of residence, income differences and politics as central anchors of modern social identity. Place of residence and geographic proximity provide tangible contexts through which individuals construct belonging and define "in-group" boundaries (Hauge, 2007; Meyners et al., 2017; Panagopoulos et al., 2017; Rijnks and Strijker, 2013; Sosa and Maoret, 2023). Similarly, socioeconomic positioning shapes cultural and world-views and influences how individuals navigate their social relations (Kraus et al., 2011; Macků et al., 2020; Moss et al., 2023; Owuamalam et al., 2017). Political identity, in turn, has become a dominant source of social categorisation, strengthening in-group attachment and reinforcing divisions with perceived out-groups (Iyengar et al., 2025; Robalo et al., 2017; Sgroi et al., 2021; van Oosten, 2025).

By combining these two dimensions, the thesis captures both contextual and self-perceived foundations of social closeness, allowing us to examine how identity-relevant cues shape cooperative and evaluative behaviour in strategic interactions.

1.4 Observability and third-party responses

Observation and evaluation are central features of organisational behaviour. Whether in public policy, government or academia, individuals routinely assess the behaviour of others, often without direct consequences for themselves. These evaluative processes can promote accountability and sustain cooperation, yet they are also shaped by identity and relatedness. People tend to be more lenient toward close others and more severe toward distant ones, highlighting the importance of examining the extent to which social proximity influences how transgressions are perceived and judged.

In the thesis, observability plays two distinct roles. Chapter 2 studies whether being observed by a socially close or distant counterpart changes own reporting under neutral incentives and independent payoffs. Chapter 3 introduces an explicit evaluative stage in which observers receive a signal about another participant's report and decide whether to impose a penalty or grant a reward. Observers' choices are non-reciprocal and non-strategic, as they affect only the counterpart's payoff, thereby isolating the social component of evaluative behaviour.

Our framework distinguishes two complementary dimensions. First, assessing whether close others are systematically rewarded more (or punished less) than distant others for equivalent signals of dishonesty. Second, analysing whether reactions to identical signals differ by social distance and socioeconomic status. This design avoids common confounds in prior work by keeping payoffs independent and separating roles between observed and observing participants.

This approach fills a gap in the literature, where reactions to observation are often incentivised through behavioural contagion, mimicry, or imitation (Dimant, 2019; Gino et al., 2009; Gueguen et al., 2009; Jordan et al., 2019), or through direct reciprocity when observers are personally affected by others' cheating (Kim and Tsvetkova, 2021; Tsvetkova and Macy, 2015). By contrast, the present framework isolates evaluative choices as expressions of social judgment rather than strategic or payoff-driven responses.

Experimental paradigms

2.1 Iterated stag-hunt game

The stag-hunt game is a classic strategic game where players choose between a risky but socially optimal choice and a safe, lower-payoff alternative. The setting in Chapter 1 implements a repeated version of this game, allowing us to test the stability of coordination under varying information visibility conditions. This structure allows for a clean test of whether trust-based traits and their disclosure influence coordination outcomes through strategic beliefs and discrimination.

Our design contributes to the literature by embedding personality disclosure into a repeated game. Whereas most stag-hunt experiments focus on payoff structures and social framing (e.g., Boone et al. (2010); Cooper (2019)), our study introduces social categorisation through personality traits as a novel coordination-relevant cue. The integration of trust-related traits and repeated interaction builds on insights from Dungan et al. (2019), showing how social cognition and trait inference affect game-theoretic choices.

2.2 Die-under-the-cup task (DUTC)

The Die-under-the-cup (DUTC) paradigm is a cornerstone in the experimental study of cheating behaviour. In this task, participants roll a die and report the outcome to determine their payoff. The anonymity of the roll provides an opportunity to lie for monetary gain, and cheating is inferred statistically by comparing aggregate reported values to the uniform distribution expected from honest reporting. The online DUTC is particularly valuable for detecting non-incentivised cheating behaviour in full anonymity, as it resembles real-world scenarios. Its flexibility allows for the manipulation of incentives, observability and framing to study the social cues the thesis is interested in exploring.

Our experimental model introduces within-subject rotation, exposing participants to socially close and socially distant counterparts. This allows to avoid confounds such as coordination or reputational incentives as in previous feedback-based DUTC studies (e.g., Kroher and Wolbring (2015)). Furthermore, unlike models that focused on collective punishment and competitive

incentives (see Benistant et al. (2021); Siniver et al. (2022)), we isolate voluntary dishonest behaviour by removing material consequences for the observer.

1.6 Chapter overview

This section provides an overview of how the thesis is structured, summarising the content of each chapter and their contributions to the overall research. Each of these chapters addresses specific research questions that jointly contribute to the objective detailed in the previous pages.

- **Chapter 1. More Predictable, Less Cooperative: The Effects of Personality Disclosure in Strategic Interaction**

  This chapter investigates how making personality information visible affects belief formation and coordination in trust-sensitive environments. Using a repeated stag-hunt game with belief elicitation, the experiment splits participants into *trusting* and *mistrusting* types and randomly assigns them to one of four information visibility regimes: no information, private information, public information and full information.

  The chapter shows that when personality traits are not disclosed, trusting and mistrusting types behave similarly. Once labels are introduced, behaviour is driven primarily by expectations rather than self-identification or type-based preferences. While personality labels shape expectations in favour of trusting counterparts, this effect fades with feedback, yet visibility improves predictability and raises strategic alignment toward the safer, inefficient equilibrium.

- **Chapter 2. How close is close enough? When social closeness backfires on honesty**

  This chapter explores the association between social closeness and cheating behaviour. Using an online Die-under-the-cup (DUTC) task, participants were asked to report private die outcomes that determined their payoffs to examine whether they differ when paired with socially close vs socially distant counterparts. The design introduced two treatments to make social closeness salient: one based on socioeconomic status (T1) and another on political alignment (T2).

  Across pooled and treatment-specific analyses, the chapter shows little systematic evidence that social closeness increases dishonest reporting. Differences in reported payoffs between close and distant pairings are small and statistically fragile. Similarly, being observed by a socially close counterpart does not reliably amplify misreporting, aside from isolated, context-dependent patterns.

- **Chapter 3. Third-party punishment and reward across socioeconomic distance**

  This chapter examines how participants respond to others' potentially dishonest behaviour, asking whether individuals punish socially distant counterparts more harshly

and reward close counterparts more generously. Across 24 rounds of an online DUTC game, the chapter introduces an active third-party response framework with participants assigned to fixed roles: die-rollers and observers, where each observer interacted in two 12-round blocks with socially close and socially distant counterparts, under one of three treatment conditions: Punishment, Reward or Mixed.

Results show that observers extend leniency toward close individuals and their responses are selective across socioeconomic levels. The chapter advances our understanding of how social closeness biases third-party judgment, highlighting that it is socially selective.

1.7 Contributions and policy implications

This thesis integrates visibility, belief formation and group affiliation within a single experimental pipeline to study how social information reshapes strategic behaviour. Across three experimental studies, it shows that social information is ambivalent: disclosure can improve coordination and predictability, but it can also distort expectations and activate bias when people judge others' behaviour.

Methodological and academic contributions. First, the thesis offers a coherent experimental framework that links information visibility, belief elicitation, behaviour and third-party responses. Chapter 1 advances the literature on information disclosure by testing multiple visibility regimes and tracing four micro-mechanisms that separate learning dynamics from one-off signalling effects—moving beyond the commonly used one-shot designs. Chapter 2 introduces a multi-scale operationalisation of social closeness, combining objective and subjective indicators. Using a within-participant, two-block observation design, this chapter isolates passive observability from strategic enforcement and finds little robust evidence that social closeness systematically increases dishonest reporting. Chapter 3 contributes an observer-level suspicion proxy and a design that isolates evaluative choice from reciprocity and payoff interdependence, and it documents selective enforcement patterns that vary with social proximity.

Policy implications. Beyond its academic relevance, the thesis presents some policy insights. First, it shows that transparency is not a panacea, as making attributes visible can improve coordination but also crystallise biases, so policies that promote disclosure should be designed with caution. Second, enforcement and monitoring systems operate within social networks, as observers may apply leniency toward socially close others, therefore anti-fraud measures that ignore social structure risk uneven application. Third, because the observed effects of social closeness are generally modest in low-stake, weakly sanctioned settings, policymakers should prioritise adjusting incentive structures and sanction mechanisms alongside any manipulation of observer identity.

Rising social fragmentation and declining institutional trust underscore the need to understand how affiliation and perceived similarity influence ethical behaviour. Integrity and monitoring initiatives often presume that individuals apply uniform judgement standards, yet the results here show that ethical judgments and sanctions can also depend on social proximity and perceived group membership. Effective policy design should therefore incorporate social dynamics—recognising that transparency, accountability, and fairness mechanisms operate within

networks of affiliation, not in isolation.

In sum, this thesis argues that dishonest behaviour is shaped as much by relational and informational context as by individual preferences. By combining rigorous experimental design with pragmatic measures of social closeness and enforcement, it delivers both a theoretical contribution to behavioural science and a practical framework for designing more effective, equitable integrity policies in fragmented and polarised environments.

Introduction générale (français)

Les interactions sociales s'appuient sur des informations concernant autrui pour orienter la confiance, la coordination et l'application des normes. Lorsque ces informations deviennent visibles, elles influencent la formation des croyances, les décisions stratégiques et l'application des normes sociales. Comprendre ces processus est central pour l'économie comportementale et la psychologie sociale, car cela met en évidence la façon dont les perceptions de similarité et d'identité peuvent favoriser la coopération tout en générant des biais. Cette thèse étudie comment la visibilité d'informations sociales reconfigure les croyances et les comportements dans des contextes stratégiques, et si ces effets dépendent de préférences pour des autres perçus comme similaires. A travers de trois études expérimentales, elle analyse comment des indices liés à l'identité, tels que les traits de personnalité, le statut socio-économique ou encore l'affiliation politique, façonnent le comportement et la formation des croyances dans des environnements stratégiques.

Motivation et questions de recherche

Apprendre l'identité des autres est crucial dans les interactions stratégiques. La dernière vague de la *World Values Survey (WVS)* rapporte que 75 % des répondants, dans la section *Capital social, confiance et appartenance organisationnelle*, estiment qu'il est nécessaire d'être « très prudent dans ses relations avec autrui ». En ventilant la confiance par groupes sociaux (par ex. voisins, connaissances, autres nationalités), aucun de ces groupes n'atteint le seuil de 25 %, à l'exception notable de la famille (77 %) (voir Haerpfer et al. (2020)). De même, l'*Enquête sur les moteurs de la confiance dans les institutions publiques* de l'OCDE (2024) fait état d'une hausse de la part des personnes déclarant une faible confiance, voire aucune confiance, dans leur gouvernement national (44 %), tendance portée par des facteurs tels que les conditions socio-économiques, le genre ou les caractéristiques démographiques (voir OECD (2024)).

Ces données illustrent la complexité des interactions où la confiance est en jeu et qui imprègnent souvent la vie sociale. Les preuves suggèrent que, pour obtenir des issues coopératives, il faut faire confiance aux autres et croire que les autres nous font confiance en retour (Hilbig et al., 2013; Rothstein and Eek, 2009). Si la plupart des individus affichent une préférence pour des partenaires dignes de confiance, la navigation des interactions quotidiennes révèle une réalité plus nuancée des dynamiques de confiance et de coopération (Barranti et al., 2016; Gächter and Schulz, 2016; Rothstein, 2011).

Comprendre comment ces dynamiques émergent, ainsi que les mécanismes qui les facilitent, demeure un objectif central de l'économie comportementale et expérimentale. Dans ce cadre, des indices de similarité — traits de personnalité partagés, milieu socio-économique ou affiliation politique — aident les individus à anticiper le comportement d'autrui. Les recherche existante indique que, lorsque l'information sur les autres est rare, les individus s'appuient sur des attentes assortatives, en supposant que les autres pensent et agissent comme eux-mêmes (Beer and Watson, 2008; Thielmann et al., 2020). Ces hypothèses servent de raccourcis cognitifs qui renforcent la prévisibilité perçue et entretiennent un sentiment de compréhension mutuelle dans l'interaction (Martin, 2015). Toutefois, ces mêmes indices peuvent aussi déclencher des biais,

de la faveur envers l'endogroupe et une application sélective des normes, créant une tension fondamentale : l'information qui facilite la coordination peut simultanément fragiliser l'équité. Lorsque l'information sur autrui est rendue visible, les croyances peuvent devenir motivées et asymétriques (Baccini and Hartmann, 2022), la coopération avec des partenaires perçus comme similaires être encouragée (Chierchia and Coricelli, 2015; Lönnqvist et al., 2021; Rubinstein and Salant, 2016), et les perceptions d'autrui se modeler à partir de croyances centrées sur soi (Cooper and Withey, 2009). Dans ce contexte, la question de recherche générale de cette thèse est double :

1. Comment la visibilité de l'information sociale affecte-t-elle les croyances, les comportements et les réactions de tiers dans des interactions stratégiques ?

2. Ces effets dépendent-ils des préférences pour des partenaires perçus comme similaires ?

La thèse aborde ces questions en développant un cadre unifié qui intègre la saillance de l'information sociale, les indices de proximité sociale et les réponses sélectives au comportement d'autrui. À travers trois chapitres expérimentaux, elle examine comment la visibilité d'informations concernant autrui — tels que les traits de personnalité, le milieu socio-économique ou l'affiliation politique — façonne la coordination, l'honnêteté et les réactions de tiers face au (non-)respect des règles. En somme, ces études visent à éclairer le double rôle de l'information sociale : à la fois mécanisme facilitant la coopération et la compréhension mutuelle, et vecteur potentiel de biais et d'indulgence envers les partenaires similaires.

Mécanismes théoriques

Cette section présente les mécanismes théoriques qui sous-tendent le cadre d'analyse et décrit la contribution de chaque chapitre aux questions de recherche générales. Chaque expérience manipule la visibilité d'informations sociales et mesure comment cela affecte la formation des croyances, le comportement et les réactions de tiers.

1.1 Divulgation d'information dans les interactions stratégiques

L'information joue un rôle central dans les interactions stratégiques. Ce que les individus savent — et savent que les autres savent — aide à aligner les attentes et à réduire l'incertitude, facilitant la coordination vers des issues mutuellement bénéfiques. Rendre une information pertinente de connaissance commune peut ainsi accélérer la convergence vers des équilibres efficaces, en particulier dans des contextes répétés où la divulgation soutient l'apprentissage et la formation de la confiance (De Freitas et al., 2019; Thomas et al., 2014; Van Huyck et al., 2018; **?**).

Cependant, la révélation d'information peut aussi agir comme signal social. Les traits visibles et les étiquettes activent la catégorisation sociale et les biais en faveur de l'endogroupe, façonnant les perceptions de similarité et de différence (Akerlof and Kranton, 2000; Tajfel and Turner, 1979). Il en résulte que la transparence peut à la fois accroître la prévisibilité et favoriser la discrimination. Par exemple, révéler une faible prosocialité peut réduire la coopération dans l'ensemble, tandis que rendre des identités de groupe de connaissance commune peut déclencher

des comportements préférentiels ou d'exclusion (Balliet et al., 2014; Bernard et al., 2016; Drouvelis and Georgantzis, 2019; Ruzzier and Woo, 2023).

Ces deux perspectives — l'information comme dispositif de coordination et l'information comme signal social — sont souvent examinées séparément. En pratique, elles coexistent et peuvent interagir. La même information qui améliore la coordination peut aussi activer des biais qui fragilisent l'équité. Comprendre l'entrelacement de ces dynamiques est particulièrement pertinent dans des environnements sensibles à la confiance, où les attentes sont fragiles et où des indices identitaires peuvent rapidement infléchir le comportement. Cette thèse aborde expérimentalement cette dualité.

Nous manipulons systématiquement ce que les participants savent d'eux-mêmes et de leurs partenaires afin d'analyser comment la visibilité de l'information façonne croyances, attentes et décisions stratégiques. Dans le **Chapitre 1**, l'information divulguée concerne des étiquettes de traits classant chaque participant comme « confiant » ou « méfiant ». La visibilité varie selon quatre conditions : *Sans information* : aucun des deux ne connaît de type ; *Information privée* : chacun ne connaît que son propre type ; *Information publique* : chacun ne connaît que le type du partenaire ; *Information complète* : chacun connaît les deux types.

Ce dispositif permet de tester si la divulgation d'information facilite la coordination en réduisant l'incertitude, ou si elle déclenche au contraire la catégorisation sociale et une coopération sélective. Les **Chapitres 2** et **3** étendent cette logique à l'information socio-économique et politique, afin de tester comment des indices liés à l'identité influencent la tricherie lors de la déclaration de résultats et les réponses de tiers face aux soupçons de tricherie.

1.2 Proximité sociale et préférence pour les semblables

Le besoin d'appartenance est une caractéristique inhérente à la condition humaine. Le partage de traits de personnalité, d'identités communes ou d'affiliations perçues accroît le sentiment de proximité qui relie les individus.

Traditionnellement, la théorie économique associe des liens forts de proximité à une confiance et une coordination plus élevées, tant dans des relations naturelles que dans des environnements de laboratoire où l'on induit des identités minimales (Bicchieri et al., 2022; Jansson, 2015; Régner and Monteil, 2007; Rubinstein and Salant, 2016). Toutefois, ces mêmes indices peuvent déplacer les standards aux frontières de groupe : la faveur envers l'endogroupe et la saillance identitaire peuvent engendrer une indulgence sélective envers les proches et une vigilance accrue envers les distants, surtout dans des environnements compétitifs ou de rareté des ressources (Bilancini et al., 2020; Chae et al., 2022; Leidner et al., 2010).

Une manifestation de ces effets est la malhonnêteté. Les individus sont plus enclins à tolérer ou à adopter des entorses aux règles qui bénéficient aux proches, et à évaluer plus positivement ces comportements (Conrads et al., 2013; Gross et al., 2018; Hoffmann, 2013; Jordan et al., 2024; Korbel, 2016; Leidner et al., 2010; Weisel and Shalvi, 2015; **?**). Des preuves en laboratoire et sur le terrain montrent que, pour des signaux comparables, les personnes sont plus indulgentes avec les proches et plus strictes avec les distants (Anvari et al., 2019; Ellemers et al., 2008; Rullo et al., 2024; Waytz et al., 2013).

La thèse traite la malhonnêteté comme un cas illustratif d'un schéma plus large où la proximité sociale façonne les attentes et reconfigure les attitudes face à l'observation de la tricherie.

Dans la suite, nous examinons ces mécanismes au moyen de conceptualisations objectives et subjectives de la proximité. Le Chapitre 2 teste si la proximité sociale (statut socio-économique et affiliation politique) modifie les attentes et le comportement de déclaration propre sous incitations neutres, gains indépendants et observation passive. Cela isole les canaux informationnels et identitaires de la réciprocité ou de la pression des pairs. Le Chapitre 3 se concentre sur la manière dont la proximité socio-économique façonne la punition et la récompense de tiers, en décomposant les décisions de tiers tout en maintenant constants les signaux statistiques de suspicion de malhonnêteté.

1.3 Modéliser la proximité sociale

La théorie de l'identité sociale soutient que l'identité d'un individu découle de la conscience de son appartenance à un groupe et de la valeur émotionnelle attachée à cette affiliation (Tajfel and Turner, 1979). La littérature montre que les traits et caractéristiques contribuent à la formation de l'identité à des degrés variables.

Si des caractéristiques visibles (par exemple le genre, la nationalité, l'origine) ont longtemps dominé les travaux expérimentaux, des recherches récentes mettent l'accent sur des indices subjectifs — croyances, statut perçu, traits de personnalité — comme fondements supplémentaires du lien et du sentiment d'appartenance. Dans cette thèse, nous modélisons la proximité sociale selon deux dimensions : le statut socio-économique (SES) et l'affiliation politique, chacune déclinée en mesure objective et subjective.

Pour le SES, la mesure objective correspond au revenu moyen de la localité de résidence du participant, tandis que la mesure subjective reflète le revenu auto-déclaré rapporté à la moyenne départementale. Pour l'affiliation politique, la mesure objective est l'orientation du parti du député élu dans la circonscription, et la mesure subjective capture les préférences individuelles vis-à-vis des principaux partis. Les indicateurs objectifs ancrent ainsi les indices identitaires dans l'environnement contextuel, plausible exogène, tandis que les indicateurs subjectifs saisissent l'identification perçue et l'alignement personnel.

Ce choix de modélisation reflète la place centrale — dans l'identité sociale contemporaine — du lieu de résidence, des écarts de revenus et de la politique. Le territoire et la proximité géographique fournissent des contextes tangibles à partir desquels se construisent l'appartenance et les frontières de l'≪ endogroupe ≫ (Hauge, 2007; Meyners et al., 2017; Panagopoulos et al., 2017; Rijnks and Strijker, 2013; Sosa and Maoret, 2023). De même, la position socio-économique structure les représentations culturelles et morales et influence la manière de naviguer les relations sociales (Kraus et al., 2011; Macků et al., 2020; Moss et al., 2023; Owuamalam et al., 2017). L'identité politique, enfin, est devenue un vecteur majeur de catégorisation, renforçant l'attachement à l'endogroupe et les divisions avec les exogroupes (Iyengar et al., 2025; Robalo et al., 2017; Sgroi et al., 2021; van Oosten, 2025). En combinant ces deux dimensions, la thèse saisit les fondations à la fois contextuelles et subjectives de la proximité sociale, ce qui permet d'examiner comment des indices identitaires orientent les comportements coopératifs et les jugements évaluatifs dans des interactions stratégiques.

1.4 Observabilité et réponses de tiers

L'observation et l'évaluation sont au cœur du fonctionnement des organisations. Dans la sphère publique, l'administration ou le monde académique, les individus apprécient régulièrement le comportement d'autrui, souvent sans conséquence matérielle pour eux-mêmes. Ces processus évaluatifs peuvent favoriser la responsabilisation et soutenir la coopération, mais ils sont aussi sensibles à l'identité et au lien social. On tend à se montrer plus indulgent envers les proches et plus strict envers les distants, d'où l'importance d'étudier dans quelle mesure la proximité sociale infléchit la perception et le jugement des transgressions.

Dans cette thèse, l'observabilité assume deux rôles distincts. Le Chapitre 2 étudie si être observé par un partenaire socialement proche ou distant modifie la déclaration propre sous incitations neutres et gains indépendants. Le Chapitre 3 introduit une phase explicitement évaluative où des observateurs reçoivent un signal sur la déclaration d'un autre participant et décident d'imposer une pénalité ou d'accorder une récompense. Les choix des observateurs sont non réciproques et non stratégiques — ils n'affectent que le gain du partenaire — ce qui isole la composante sociale du jugement.

Notre cadre distingue deux dimensions complémentaires. Premièrement, vérifier si, à signal de malhonnêteté équivalent, les proches sont systématiquement plus récompensés (ou moins sanctionnés) que les distants. Deuxièmement, analyser si les réactions à des signaux identiques diffèrent selon la distance sociale et le statut socio-économique. Ce dispositif évite des confusions fréquentes dans la littérature en maintenant des gains indépendants et en séparant les rôles d'observé et d'observateur.

Cette approche comble un manque des travaux existants, où les réactions à l'observation sont souvent étudiées via la contagion comportementale, le mimétisme ou l'imitation (Dimant, 2019; Gino et al., 2009; Gueguen et al., 2009; Jordan et al., 2019), ou via la réciprocité directe lorsque l'observateur subit personnellement la tricherie d'autrui (Kim and Tsvetkova, 2021; Tsvetkova and Macy, 2015). À l'inverse, notre cadre isole les choix évaluatifs comme expressions d'un jugement social plutôt que comme réponses stratégiques ou dictées par le gain.

Paradigmes expérimentaux

2.1 Jeu de type *stag-hunt* répété

Le jeu de « stag-hunt » est un jeu économique classique où les joueurs choisissent entre une option risquée mais socialement optimale et une alternative sûre à gain plus faible. Le cadre du chapitre 1 met en place une version répétée, permettant de tester la stabilité de la coordination sous différents régimes de visibilité de l'information. Cette structure offre un test net de l'influence des traits liés à la confiance et de leur divulgation sur les issues de coordination via les croyances stratégiques et la discrimination.

Notre contribution consiste à intégrer la revelation de traits de personnalité dans un jeu répété. Alors que la plupart des expériences stag-hunt se concentrent sur les structures de gains et le cadrage social (par ex. Boone et al. (2010) ; Cooper (2019)), notre étude introduit une catégorisation sociale fondée sur des traits de personnalité comme nouvel indice pertinent pour la coordination. L'articulation entre traits liés à la confiance et interaction répétée prolonge les

enseignements de Dungan et al. (2019), en montrant comment la cognition sociale et l'inférence de traits affectent les choix en situation stratégique.

2.2 Tâche « Die-under-the-cup » (DUTC)

Le paradigme DUTC est un pilier de l'étude expérimentale de la tricherie. Les participants lancent un dé et déclarent le résultat qui détermine leur gain ; l'anonymat du lancer offre l'opportunité de mentir, et la tricherie est inférée statistiquement en comparant les déclarations agrégées à la distribution uniforme attendue sous honnêteté. La version en ligne est particulièrement utile pour détecter des comportements de tricherie non incités en pleine anonymat, proches de situations réelles. Sa flexibilité permet de manipuler incitations, observabilité et cadrage pour explorer les indices sociaux étudiés dans la thèse.

Notre dispositif introduit une rotation intra-sujet, exposant les participants à des partenaires socialement proches et distants. Cela permet d'éviter des confusions comme des incitations de coordination ou de réputation, fréquentes dans des DUTC avec rétroaction (voir par ex. Kroher and Wolbring (2015)). De plus, à la différence de modèles axés sur la punition collective ou des incitations compétitives (voir Benistant et al. (2021); Siniver et al. (2022)), nous isolons la malhonnêteté volontaire en supprimant toute conséquence matérielle pour l'observateur.

1.6 Aperçu des chapitres

Cette section présente l'organisation de la thèse, en résumant le contenu de chaque chapitre et leur contribution à l'ensemble du projet de recherche. Chacun des chapitres répond à des questions spécifiques qui, conjointement, nourrissent l'objectif de recherche détaillé précédemment.

- **Chapitre 1. Plus prévisible, moins coopératif : les effets de la divulgation des traits de personnalité dans l'interaction stratégique**

  Ce chapitre analyse comment la visibilité d'informations de personnalité influe sur la formation des croyances et la coordination dans des environnements sensibles à la confiance. À l'aide d'un jeu de chasse au cerf répété avec éllicitation des croyances, l'expérience classe les participants en types *confiants* et *méfiants* et les assigne aléatoirement à l'un de quatre régimes de visibilité de l'information : sans information, information privée, information publique et information complète.

  Le chapitre montre que, lorsque les traits de personnalité ne sont pas divulgués, les types confiants et méfiants se comportent de manière similaire. Une fois les étiquettes introduites, le comportement est principalement guidé par les attentes plutôt que par l'auto-identification ou des préférences de type. Les étiquettes biaisent d'abord les croyances de premier ordre en faveur des partenaires confiants, mais cet effet s'estompe avec le retour d'information. La visibilité améliore la prévisibilité et accroît l'alignement stratégique vers l'équilibre plus sûr mais inefficace. En somme, le chapitre 1 met en évidence le double rôle de l'information sociale : elle peut à la fois favoriser l'alignement et, via des distorsions de croyances, compromettre la coopération efficiente.

- **Chapitre 2. Jusqu'où peut aller la proximité sociale ? Quand la proximité sociale se retourne contre l'honnêteté**

  Ce chapitre explore le lien entre proximité sociale et comportement de tricherie. Dans une tâche en ligne de type Die-under-the-cup (DUTC), les participants déclarent en privé des résultats de dé déterminant leurs gains, ce qui permet d'examiner si les déclarations diffèrent selon que l'interaction a lieu avec des partenaires socialement proches ou socialement distants. Le dispositif introduit deux traitements rendant saillante la proximité sociale : l'un fondé sur le statut socio-économique (T1) et l'autre sur l'alignement politique (T2).

  Dans des analyses globales et spécifiques par traitement, le chapitre met en évidence peu d'éléments systématiques montrant que la proximité sociale accroît la malhonnêteté. Les écarts de gains déclarés entre paires proches et distantes sont faibles et statistiquement fragiles. De même, être observé par un partenaire socialement proche n'amplifie pas de manière fiable les fausses déclarations, en dehors de motifs ponctuels dépendants du contexte.

- **Chapitre 3. Sanction et récompense par des tiers selon la distance socio-économique**

  Ce chapitre étudie la façon dont les participants réagissent au comportement potentiellement malhonnête d'autrui, en se demandant si les individus sanctionnent plus sévèrement les partenaires socialement distants et récompensent plus généreusement les proches. Sur 24 manches d'un DUTC en ligne, le chapitre introduit un cadre d'intervention active de tiers, avec des rôles fixes : lanceurs de dé et observateurs. Chaque observateur interagit dans deux blocs de 12 manches avec des partenaires socialement proches et socialement distants, sous l'un de trois traitements : Punition, Récompense ou Mixte.

  Les résultats montrent que les observateurs font preuve d'indulgence envers les proches et que leurs réponses sont sélectives selon le niveau socio-économique. Le chapitre met en lumière la manière dont la proximité sociale biaise le jugement des tiers, soulignant que l'évaluation est socialement sélective.

1.7 Contributions et implications pour les politiques publiques

Cette thèse intègre visibilité de l'information, formation des croyances et appartenance de groupe au sein d'un même protocole expérimental afin d'étudier comment l'information sociale reconfigure les comportements stratégiques. À travers trois études expérimentales, elle montre que l'information sociale est ambivalente : la divulgation peut améliorer la coordination et la prévisibilité, mais elle peut aussi déformer les attentes et activer des biais dans le jugement du comportement d'autrui.

Apports méthodologiques et académiques. Premièrement, la thèse propose un cadre expérimental cohérent reliant visibilité de l'information, élucidation des croyances, comportement et réponses de tiers. Le chapitre 1 fait progresser la littérature sur la divulgation d'attributs en testant plusieurs régimes de visibilité et en retraçant quatre micro-mécanismes permettant

de distinguer la dynamique d'apprentissage des effets de signalisation ponctuels — dépassant ainsi les cadres usuellement fondés sur des designs one-shot. Le chapitre 2 introduit une opérationnalisation multi-échelle de la proximité sociale, combinant indicateurs objectifs et subjectifs. Grâce à un dispositif d'observation intra-sujets en deux blocs, ce chapitre isole l'observabilité passive de l'application stratégique de sanctions et montre peu d'éléments robustes en faveur d'un effet systématique de la proximité sociale sur la tricherie déclarée. Le chapitre 3 apporte un dispositif méthodologique supplémentaire : un indice de suspicion au niveau de l'observateur et un design qui sépare le choix évaluatif (punir/récompenser) de la réciprocité et de l'interdépendance des gains ; il documente des schémas d'application sélective des sanctions qui varient selon la proximité sociale.

Implications pour les politiques publiques. Au-delà de son intérêt académique, la thèse délivre plusieurs enseignements pour les politiques publiques. Premièrement, la transparence n'est pas une panacée : rendre des attributs visibles peut améliorer la coordination mais aussi cristalliser des biais, de sorte que les politiques favorisant la divulgation doivent être conçues avec prudence. Deuxièmement, les systèmes de contrôle et d'application (monitoring, sanctions) fonctionnent au sein de réseaux sociaux : les observateurs peuvent faire preuve de clémence envers des proches socialement affiliés, et des dispositifs anti-fraude qui ignorent la structure sociale risquent une application inégale. Troisièmement, comme les effets de la proximité sociale sont généralement modestes dans des contextes faiblement sanctionnés et à faible enjeu, les décideurs devraient prioriser l'ajustement des incitations et de la sévérité des sanctions en complément de toute manipulation de l'identité de l'observateur.

La montée de la fragmentation sociale et la baisse de la confiance institutionnelle renforcent la nécessité de comprendre comment l'affiliation et la similitude perçue influencent le comportement éthique. Les dispositifs d'intégrité et de contrôle supposent souvent que les individus appliquent des normes morales uniformes ; nos résultats montrent que les jugements éthiques et les sanctions peuvent aussi dépendre de la proximité sociale et de l'appartenance perçue à un groupe. La conception efficace des politiques doit donc intégrer les dynamiques sociales — en reconnaissant que transparence, responsabilité et équité s'exercent au sein de réseaux d'affiliation et non en dehors de tout contexte social.

En somme, cette thèse soutient que la malhonnêteté ne peut être pleinement comprise sans tenir compte des contextes relationnels et informationnels dans lesquels les décisions sont prises. En combinant un design expérimental rigoureux et des mesures pragmatiques de la proximité sociale et de l'application des normes, elle apporte à la fois une contribution théorique aux sciences du comportement et un cadre opérationnel utile pour concevoir des politiques d'intégrité plus efficaces et plus équitables dans des sociétés fragmentées et polarisées.

Chapter 1. More Predictable, Less Cooperative: The Effects of Personality Disclosure in Strategic Interaction

Irving Argaez Corona[a], Béatrice Boulu-Reshef[b], Jean-Christophe Vergnaud[a,c]

[a]Centre d'Économie de la Sorbonne (CES), Université Paris 1 Panthéon-Sorbonne, France
[b]CY Cergy Paris Université (THEMA), France
[c]Centre National de la Recherche Scientifique (CNRS), France

Abstract

This study investigates how the disclosure of personality traits affects strategic behaviour in trust-sensitive environments. Specifically, we examine whether making information about personality traits visible—either about oneself, one's counterpart, or both—modifies coordination outcomes. In a repeated stag-hunt game, 192 participants were classified as either trusting or mistrusting types and randomly assigned to one of four information conditions (no information, private, public, or full visibility). We elicited first and second-order beliefs across 48 rounds to analyse how trait visibility shapes expectations and behaviour through four mechanisms: (1) self-identification, (2) preference-based discrimination, (3) first-order belief bias, and (4) second-order belief pessimism. Our results show that when traits are not disclosed, personality has no behavioural effect: trusting and mistrusting types look alike. Once labels appear, behaviour is driven by expectations, not by self-identification or type-based preferences. We observe that labels shift beliefs toward trusting counterparts, but this fades with feedback. Moreover, information increases predictability and strategic alignment, yet can nudge play toward the safer, inefficient equilibrium—so coordination rises even as cooperation can fall.

This project has benefitted from funding from the 2022 allocation campaign of the *Centre d'Économie de la Sorbonne (CES)* of Université Paris 1 Panthéon-Sorbonne. The paper received ethics approval from the Institutional Review Board of Paris School of Economics (decision 2022-013). The pre-registered research protocol can be found in this link for peer-review purposes.

We warmly thank **Maxim Frolov**, computer engineer at the *Laboratoire d'Economie Expérimentale de Paris (LEEP)*, for his technical support in programming the experiment and his invaluable help in organising the experimental sessions.

Résumé

Cette étude examine comment la divulgation de traits de personnalité affecte le comportement stratégique dans des environnements sensibles à la confiance. Plus précisément, nous analysons si rendre visibles des informations de personnalité — sur soi-même, sur son partenaire, ou sur les deux — modifie les résultats de coordination. Dans un jeu répété du type *stag-hunt*, 192 participants ont été classés comme « confiants » ou « méfiants » et assignés aléatoirement à l'un de quatre régimes d'information (aucune information, information privée, information publique ou visibilité totale). Nous avons recueilli des croyances de premier et de second ordre sur 48 manches afin d'étudier comment la visibilité des traits façonne attentes et comportements via quatre mécanismes : (1) auto-identification, (2) discrimination fondée sur le type, (3) biais de croyances de premier ordre et (4) pessimisme de croyances de second ordre. Nos résultats montrent que, en l'absence de divulgation, la personnalité n'a pas d'effet comportemental : les types confiants et méfiants se ressemblent. Une fois les étiquettes visibles, le comportement est principalement guidé par les attentes, non par l'auto-identification ni par des préférences de type. Nous observons que les étiquettes déplacent temporairement les croyances en faveur des partenaires « confiants », mais cet effet s'estompe avec le feedback. Par ailleurs, l'information accroît la prévisibilité et l'alignement stratégique, tout en pouvant orienter le jeu vers l'équilibre plus sûr mais inefficace — la coordination augmente alors même que la coopération peut diminuer.

1.1 Introduction

Accurately anticipating others' behaviour is central to successful coordination. Work in economics and social psychology increasingly converges on the importance of expectations and social preferences in guiding cooperative choices, while acknowledging that individuals often rely on limited cues to form those expectations (Bénabou and Tirole, 2016; Castro Santa et al., 2018; Gries et al., 2022).

Within this broader view, information about other people's personality traits is key to anticipate how they will behave in strategic interactions. Thus far, the literature studying personality traits and the effects of their disclosure has produced more nuanced than uniform results. On one hand, revealing specific information about individuals' traits can reduce strategic uncertainty, improve expectations and support coordination (Bose and Sgroi, 2022; Chierchia and Coricelli, 2015; Hughes et al., 2020; Rubinstein and Salant, 2016; Thomas et al., 2014); while on the other, making personality trait information readily available can foster in-group favouritism, encourage avoidance of non-preferred personalities and reinforce negative stereotypes that are detrimental for coordination (Ahloy and Hamman, 2019; Balliet et al., 2014; Drouvelis and Georgantzis, 2019). These patterns suggest that the effects of information are heterogeneous: which information is disclosed, to who and whether disclosure happens in full or partial visibility determines expectations and, in turn, how strategic coordination unfolds.

This paper examines how disclosing personality information reshapes strategic coordination in a trust-sensitive environment. We present a unified framework that focuses on both, the benefits of information disclosure (e.g., reduced uncertainty, improved matching) and on its downsides (e.g., stereotype-driven exclusion, belief distortions) across four information regimes: no information (baseline), private information (only self label visible), public information (only counterpart's label visible), and full information (both labels visible).

We recruited 192 participants, categorised them as *trusting* or *mistrusting* using a personality questionnaire, and measured their risk attitudes and *ex-ante* beliefs before they played a 48-round stag-hunt game with belief elicitation in every round. In the stag-hunt, participants face two choices: cooperation A, the socially efficient but riskier action, or non-cooperation B, the inefficient but safe option. Therefore, four possible outcomes can emerge from every interaction, with two of these reflecting coordination (i.e., both choosing the same action A–A or B–B). The stag-hunt's multiple equilibria makes it well-suited to study how information and beliefs shape strategic alignment under uncertainty.

To understand how trait disclosure influences decisions, we test four mechanisms that reflect distinct pathways from visibility to beliefs and behaviour. **(1) Self-identification**: tests whether making labels salient prompts people to internalise their assigned label and behave in line with its stereotype. **(2) type-based discrimination**: tests whether making the counterpart's label visible, prompts participants to favour "trusting" counterparts or prefer same-type counterparts (homophily); **(3) first-order belief bias (FOB)**: tests if label visibility can skew expectations about others' cooperation (e.g., optimism toward trusting types) even when actual behaviour does not differ; and **(4) second-order belief pessimism (SOB)**: tests if when one's own label is visible, individuals anticipate being seen as uncooperative (especially if labelled mistrusting), hence lowering their expectations about how they're perceived, and behaving more

cautiously even without explicit rejection.

The objective is to test the dual role of information in coordination: it can raise strategic alignment by reducing uncertainty and improving matching, yet can hinder cooperation by activating stereotypes, distorting beliefs, and encouraging exclusion. Our contribution is to pinpoint and test the mechanisms through which trait disclosure shifts beliefs and social responses, showing how it can both support and undermine coordination, steer play toward efficient or inefficient equilibria.

Our results show that when traits are not disclosed, personality has no predictive power for behaviour, as trusting and mistrusting types do not exhibit different cooperation/coordination rates. However, once labels are introduced, behaviour is driven primarily by expectations rather than by disclosed information on self or others. We find no direct behavioural conformity when self label is visible (Mechanism 1) and no robust evidence of type-based discrimination or homophily in cooperative choices (Mechanism 2). By contrast, Mechanism 3 supports label-driven shifts in beliefs, as they favour trusting counterparts when labels are revealed. This effect fades with feedback and game repetition, mainly due to a learning effect across interactions. We also find limited support for Mechanism 4, as while SOB move strongly with the counterpart's expectations (FOB), self-relevant information induces only a small, visibility-contingent adjustment in pessimistic expectations.

In sum, our study reveals that information disclosure improves predictability and fosters coordination, however, it is beliefs and not preference-based sorting that drive these changes. Our results suggest that information disclosure reduces uncertainty, but can steer choices toward the safer and inefficient equilibria. Findings reveal that, in order to reduce non-coordination, limited information disclosure or early private feedback can improve expectations; while disclosing others' recent history of cooperative behaviour is more relevant than disclosing their personality labels to foster cooperation.

The paper proceeds as follows: Section 1.2 reports the theoretical background, related literature and where our study fits. Section 1.3 presents the experimental design and procedures. Section 1.4 presents the descriptive statistics and 1.5 reports the results. Section 1.6 concludes.

1.2 Theoretical background and related literature

Strategic coordination relies on individuals' ability to form expectations about others' behaviour. In coordination environments such as the stag-hunt game, achieving the efficient outcome requires confidence that one's counterpart will make the same cooperative choice. Although there is no opportunity for exploitation—as in standard trust games—the risk of non-coordination makes success contingent on beliefs about others' intentions.

In such settings, providing information about others can enhance coordination by reducing strategic uncertainty, but may also distort beliefs and activate stereotype-driven expectations. This distinction is central to our approach. The following section reviews research on information disclosure, coordination, and personality-based expectations, and outlines four mechanisms through which personality trait visibility may shape beliefs and behaviour in coordination games.

1.2.1 Information and coordination in strategic settings

Information is critical for trust-based decision-making. Following Rubinstein (1989), who was among the first to highlight the crucial role of full information disclosure to improve coordination, a large stream of literature shows that when relevant information becomes common knowledge strategic alignment improves (Abeler et al., 2019; Boone et al., 2010; Devetag and Ortmann, 2007; Giorgetta et al., 2021; Larrouy and Lecouteux, 2017). In games with multiple equilibria, even small changes in what players know about themselves or others can shift beliefs and behaviour (Charness, 2000; Thomas et al., 2014). Experimental studies have shown that increasing transparency—through public signals, pre-play communication, reputation or disclosed feedback—facilitates coordination by making actions more predictable and reducing strategic uncertainty (De Freitas et al., 2019; Thomas et al., 2014; Van Huyck et al., 2018; **?**).

In repeated settings, information can accelerate convergence to efficient equilibria by reinforcing mutual expectations and supporting learning (Kreps, 1992). However, when the disclosed information concerns social categories, personality traits, or group identities, its effects become more ambiguous. Social psychology has long shown that visible labels can activate stereotype-consistent behaviour, in-group biases, and distorted expectations (Akerlof and Kranton, 1997; Tajfel and Turner, 1979). In strategic environments, these distortions may lead to belief-driven exclusion or coordination breakdowns, even in the absence of objective behavioural differences (Acedo-Carmona and Gomila, 2014; Balliet et al., 2014; Bernard et al., 2016; Ruzzier and Woo, 2023). Thus, disclosure can both coordinate and polarise decisions, as it raises predictability but may also activate identity and image concerns that steer behaviours.

These two perspectives-information as a coordination device, and information as a social signal—are often studied in isolation. However, they may interact as the same trait disclosure that improves predictability may also trigger the selectivity that undermines cooperation. Understanding how these dynamics intertwine is especially relevant in trust-sensitive environments, where expectations are fragile and identity cues can quickly shape behaviour.

1.2.2 Personality traits and trust in economic behaviour

Personality traits have long been studied as determinants of trust and cooperation in behavioural economics and experimental games. Individual differences in dispositional trust—defined as a general tendency to expect benevolence from others—have been found to correlate with trusting behaviour in a variety of contexts, including trust games, public goods games, and repeated prisoner's dilemmas (Fehr et al., 2008). Measures such as the *Trust* subscale from the Big Five Inventory or dedicated trust questionnaires have been used to classify participants as "high trust" or "low trust" types, with modest but consistent predictive power.

The behavioural relevance of these traits, however, is not uniform. Several studies report limited predictive power when information about personality is absent (Chierchia and Coricelli, 2015; Currarini and Mengel, 2016), suggesting that dispositional trust may only matter when it becomes salient—either through feedback, framing, or explicit disclosure. Moreover, the effects of trait-based classification may go beyond preferences and reflect more complex cognitive or social dynamics, such as stereotype activation or belief updating.

In our setting, we classify participants exogenously as trusting or mistrusting types using a

hybrid version of two standard personality questionnaires-the NEO-PI inventory and HEXACO-focusing on the trait of agreeableness, as it is closely linked to trust, cooperation and prosocial behaviour (Rustichini et al., 2016; Thomas et al., 2016) and is one of the most extensively documented traits in game-related literature (Fagbenro, 2019; McFerran et al., 2010; Proto et al., 2014; Rustichini et al., 2016; Volk et al., 2011). The research suggests that individuals with high levels of agreeableness are more prone to cooperation (see Ahloy and Hamman (2019); Crowe et al. (2018); Drouvelis and Georgantzis (2019); Proto et al. (2014); Stahl and Huyck (2002)), but that cooperation often depends on whether individuals are aware of these traits in them or in their counterparts (Thomas et al., 2014; Zhao and Smillie, 2015). For instance, Jansson and Eriksson (2015) explored the role of trust labels in promoting cooperation, finding that mutual knowledge of trusting traits increases coordination into the same outcome.

While these findings indicate that personality traits can play a role in strategic settings, our experimental design does not assume that such traits are predictive of behaviour in the absence of contextual salience. The binary labelling we implement is not intended to capture stable behavioural types, but to serve as a socially interpretable signal.

1.2.3 Trait visibility and cognitive mechanisms in coordination

Beyond its informational value, trait visibility can alter coordination through a range of cognitive and social mechanisms. When individuals are categorised into personality types such as trusting type (TT) or mistrusting type (MT), these labels may shape both self-perception and expectations about others. In what follows, we outline four distinct mechanisms through which trait information may influence beliefs and behaviour in strategic settings.

Self-identification. When individuals are assigned a trait label such as being a "trusting" or "mistrusting" type, they may internalise the categorisation and align their behaviour accordingly, even when this information remains private. This mechanism, which we refer to as self-identification, echoes findings in identity economics and behavioural psychology, where individuals conform to socially salient roles or stereotypes (Drouvelis and Georgantzis, 2019; Falk and Zimmermann, 2024). In a coordination context, this may lead mistrusting types to act more cautiously or less cooperatively than they would otherwise, not because of intrinsic preferences, but because they see themselves through the lens of the assigned trait.

Type-based discrimination. When participants are informed of their counterpart's personality type, they may condition their strategy on the perceived desirability of that type. Specifically, trusting types may be seen as safer or more predictable cooperation partners, while mistrusting types may be avoided. We refer to this behaviour as type-based discrimination. This mechanism does not require any objective behavioural difference across types but arises from how individuals evaluate social labels in strategic settings (Balliet et al., 2014).

Type-based discrimination may take the form of out-group avoidance, meaning a tendency to favour those labelled as trusting over those labelled as mistrusting, but also of in-group preference or homophily, where individuals display a greater willingness to cooperate with others who share their own type label (Charroin et al., 2021; Currarini and Mengel, 2016). Both forms may reflect expectations of similarity, perceived compatibility, or social alignment, and can shape cooperation even in the absence of partner choice.

First-order belief bias. Even when individuals do not directly discriminate in behaviour, knowing the type of a counterpart may lead them to form biased expectations about that person's likelihood of cooperation. This distortion reflects belief-based discrimination (Rustichini et al., 2016; Thomas et al., 2016), where perceived trustworthiness is inferred from type labels rather than observed behaviour. Such effects have been shown to influence economic decisions in trust and allocation games, particularly when cognitive shortcuts or stereotypes are activated.

Second-order belief pessimism. Finally, when participants know that their own trait is visible to their counterpart, they may adjust their expectations based on how they believe they are perceived. Mistrusting types, in particular, may anticipate being seen as less cooperative and revise their beliefs downward accordingly. We refer to this as second-order belief pessimism. This mechanism is consistent with theories of anticipated rejection and stereotype threat, where individuals lower their strategic expectations out of concern for how they are viewed (Thomas et al., 2014)).

In sum, these mechanisms suggest that trait visibility can reshape cooperation through both direct behavioural responses and recursive belief formation.

1.2.4 Contribution and hypotheses

This study contributes to the understanding of how trait visibility affects strategic coordination by combining insights from coordination theory, social identity, and belief formation. While previous work has shown that personality traits can shape behaviour in economic games, and that information can alter coordination dynamics, few studies have explored how the disclosure of personality labels interacts with belief structures in a systematic and controlled environment. Our design leverages a repeated stag-hunt game with 192 participants assigned to one of four information treatments:

- **No information**: neither participant knows any type label;
- **Private information**: participants only know their own label;
- **Public information**: participants only know their counterpart's label;
- **Full information**: both participants know both labels.

Participants are categorised exogenously as either trusting or mistrusting types based on a standard personality questionnaire focused on agreeableness. In each round, we elicit both first-order beliefs (expectations about the partner's behaviour) and second-order beliefs (expectations about how one is perceived), enabling us to capture the full structure of beliefs that underpin coordination decisions. Overall, our contribution is both conceptual and methodological: we offer a framework to study the self-fulfilling nature of trait information in strategic environments, and an empirical strategy to disentangle its behavioural and cognitive consequences.

1.3 Experimental design and procedures

Overview of the sequence of tasks

Each participant engaged in a series of six tasks: **(1)** a Holt and Laury (2002) risk-attitude task; **(2)** an incentivised baseline round of a stag-hunt game with first-order and second-order belief elicitation with no feedback; **(3)** a personality questionnaire adapted from HEXACO and the NEO-PI to categorise them into *trusting type (TT)* and *mistrusting type (MT)*; **(4)** a non-incentivised *ex ante* questionnaire to (a) predict their own personality type, (b) elicit first-order beliefs of a mistrusting type, (c) elicit first-order beliefs of a trusting type, (d) elicit second-order beliefs of a mistrusting type, and (e) elicit second-order beliefs of a trusting type, prior to this task they were presented with a brief explanation of the traits attached to a trusting and a mistrusting type (see Table 1.1); **(5)** an incentivised 48-round stag-hunt game with first and second-order belief elicitation and feedback in one of the four information regimes; and **(6)** an *ex post* questionnaire to (a) register their own assumed personality type, (b) their first-order beliefs of a mistrusting type, (c) their first-order beliefs of a trusting type, (d) their second-order beliefs of a mistrusting type, and (e) their second-order beliefs of a trusting type. For an overview of the experiment's instructions, refer to Subsection 1.7 in the Appendix.

Task 1. Risk-attitude task.

The participants' risk attitudes were assessed through the Holt and Laury (2002) procedure. In this task, participants were presented with a table featuring paired lottery choices: Option A offered a lottery with a high payoff of \$2.00 and a low payoff of \$1.60, while Option B presented a lottery with a high payoff of \$3.85 and a low payoff of \$0.10. The probabilities of the high and low payoffs varied across lines, starting with 1/10 and 9/10 on the first line and progressing to 10/10 and 0/10 on the last line. Choosing Option B when the probability of the high payoff was low represented a risky decision, and as probabilities changed, the expected value of Option B over Option A increased. The measurement of risk attitude is based on the number of safe choices made.

Task 2. One-shot baseline round.

Each participant was randomly matched with another participant in the session to complete a one-shot, incentivised baseline round of the stag-hunt game. The stag-hunt is a strategic game with two options: A (cooperation) or B (no cooperation), and four possible outcomes. To avoid framing effects, labels are neutrally referred to as "Option A" and "Option B" instead of "cooperation" or "no-cooperation" across the experiment (see Figure 1.1). To begin the task, participants were informed of the operationalisation of the stag-hunt and proceeded to play the baseline round in the following sequence:

1. Report their first-order beliefs (FOB) on their perceived probability that the counterpart would choose "A" rather than "B".

2. Report their second-order beliefs (SOB) on their perceived probability about the counterpart's beliefs that self would choose "A" rather than "B".

3. Make their choice in the stag-hunt matrix.

To ensure that belief elicitation was understood and incentive-compatible, beliefs were reported in a single slider with endpoints labelled "A" and "B" and two percentages summed to 100. As participants moved the slider, they could visualise how reporting a probability closer to the realised outcome would yield higher earnings, thereby reducing misunderstanding or random reporting. For SOB elicitation, the interface additionally included a dynamic graphical display showing the expected payoff as a function of the reported probability, based on the Quadratic Scoring Rule (QSR) (see Figure 1.2). This feature was introduced to facilitate comprehension, as reasoning about second-order beliefs is less intuitive than predicting their direct action. Both sliders prevented continuation to the next page until participants had moved them at least once.

Figure 1.1: Stag-hunt game matrix

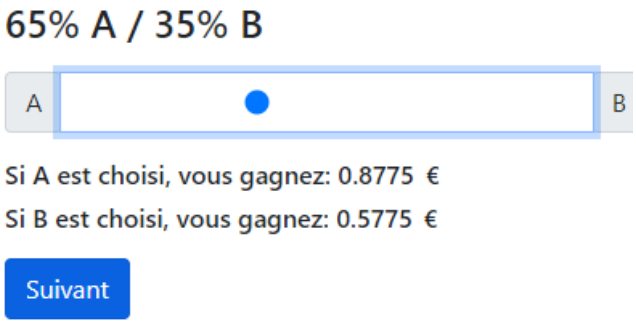
		L'autre participant	
		A	**B**
Vous	Je choisis A	6, 6	-3, 3
	Je choisis B	3, -3	3, 3

Beliefs were incentivised with a QSR to allow us to gauge the accuracy of participants' predictions: $\text{QSR} = 1 - (1 - p)^2$, where p is the reported probability. For instance, an accurate report of 70% provides a score of $1 - (1 - 0.70)^2 = 0.91$. The interface displayed the points (earnings) corresponding to each reported probability under both possible outcomes ("A" or "B"). Belief elicitation followed this procedure throughout the experiment to ensure truthful reporting, replicating the findings of Artinger et al. (2010); Greenberg and Org (2018); Hoffmann (2013); Hyndman et al. (2013); Offerman et al. (2009).

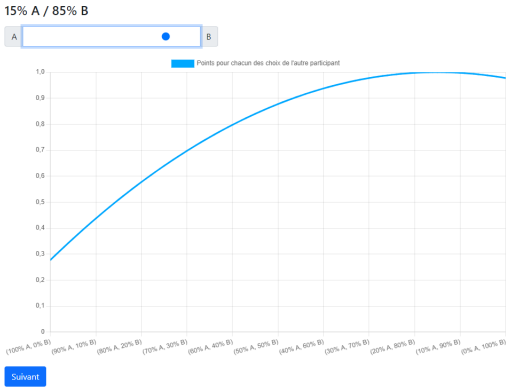
Task 3. Participant categorisation.

We implemented a hybrid personality questionnaire combining HEXACO (Ashton and Lee, 2007) and the NEO personality inventories (Costa, 1992) to categorise participants in two personality types: trusting and mistrusting types. The hybridisation of these inventories responds to their complementary measurement of the *Agreeableness* trait: HEXACO emphasises facets such as *Flexibility*, *Forgiveness* and *Gentleness*, which capture tolerance and conflict-avoidance tendencies, while NEO-PI focuses more directly on the *Trust* facet, reflecting expectations about others' honesty and reliability. Therefore, combining both scales allowed us to obtain a broader and more balanced measure of dispositional trust, integrating attitudinal components relevant for cooperation and coordination.

Figure 1.2: FOB and SOB elicitation sliders.



(a) First-order belief elicitation slider



(b) Second-order belief elicitation slider

In the questionnaire, participants ranked their level of accordance with 20 statements on a five-point *Likert* scale (1 = strongly disagree, 5 = strongly agree), with 9 of the statements reverse-scored (full questionnaire in subsection 1.12 in the Appendix). The 20 statements correspond to the Agreeableness facets of *Flexibility*, *Forgiveness*, and *Gentleness* from HEXACO and the *Trust* facet from NEO-PI, and included items such as "I fundamentally believe that most people are well-intentioned" and "If someone has already cheated on me, I would be forever mistrustful."

We used *Agreeableness* because it is the most indicative of trust-related traits and cooperative tendencies (Crowe et al., 2018; Proto et al., 2014; Rustichini et al., 2016). Both NEO (Costa and McCrae, 2008) and HEXACO (Ashton and Lee, 2008) have proven reliable predictors of traits associated with prosociality, coordination, and ethical decision-making (Coyne and Bartram, 2002; Fagbenro, 2019; Hilbig et al., 2015; John et al., 2008; Kajonius and Dåderman, 2014; Proto et al., 2014; Ruch et al., 2017). Their combination thus provided a more precise operationalisation of the trust-related personality construct relevant to our coordination framework.

We treat these scores as indicators of trusting traits for two reasons. First, they reflect participants' general tendency to expect others to act cooperatively, a psychological predisposition that shapes how signals are interpreted and how beliefs are formed in strategic settings (i.e., general trust). Second, our experimental manipulation makes trait labels visible, so it is the awareness of a counterpart being trusting or mistrusting that can alter expectations and strategic responses. Overall, modelling trusting traits aligns the empirical measure with the theoretical mechanism: how a dispositional orientation toward others conditions the translation of trait information into interpersonal beliefs and, ultimately, cooperative choices.

Categorisation was performed using the session-specific median of participants' scores: those above the median were classified as *Trusting personalities*, and those below as *Mistrusting personalities*. No feedback or payoff was given for this task.

Task 4. *Ex ante* belief elicitation questionnaire (non-incentivised).

We elicit *ex ante* beliefs on (1) participants' prediction of their own personality type, (2) first-order beliefs with respect to a mistrusting type, (3) first-order beliefs with respect to a trusting type, (4) second-order beliefs with respect to a mistrusting type, and (5) second-order beliefs with respect to a trusting type. Given that this task is non-incentivised, there is no payoff involved. At this stage, we provide participants with the description for the trusting and mistrusting labels, as reported in Table 1.1.

Table 1.1: Personality type description.

Trusting personalities	Mistrusting personalities
Trusting personalities are those who can easily trust others and believe others will trust them.	Mistrusting personalities are those who find it hard to trust others and do not believe others will trust them.

Task 5. Repeated stag-hunt game (incentivised).

We formed cohorts of 6 participants with 3 categorised as trusting and 3 as mistrusting, rotating in dyads within the same cohort across 48 rounds. Each round followed this sequence:

1. Elicited participants' first-order beliefs (FOB) asking: *You are a trusting/mistrusting personality. The other participant in this period is a trusting/mistrusting personality. What are the percentages of chance that the other participant will choose A and B in this round?.*

2. Elicited participants' second-order beliefs (SOB) asking: *The other participant is a trusting/mistrusting personality and has just estimated the probability you gave to the probability percentages between A and B. What answer do you think they gave?.*

3. Make their choice in the stag-hunt matrix (see Figure 1.1).

The stag-hunt matrix allows us to test whether disclosure helps players reach strategic coordination or, conversely, leads to a breakdown of cooperation when individuals react to the revealed label of their counterpart.

After each round, participants received feedback on (1) their decision and their counterpart's decision in the stag-hunt, (2) the level of accuracy on their FOB and SOB and (3) their earnings. Payoffs from the stag-hunt game were expressed in *Experimental Currency Units (ECUs)* and were accumulated throughout the 48 rounds of the repeated game at a rate of 1 ECU = 0.04€. This allows us to keep all rounds incentive-relevant, ensuring participants respond meaningfully to each label pairing.

As for belief elicitation, only one round of FOB and one round of SOB in the incentivised tasks (i.e., Tasks 2 and 5) were randomly drawn for payment directly in euros. Total payoffs were then the sum of the accumulated ECUs in the 48 rounds of the repeated stag-hunt, a randomly drawn round of elicited FOB, a randomly drawn round of elicited SOB, gains from the Holt & Laury risk attitude task and a 5€ show up fee.

Information regimes.

Information is disclosed across four regimes: No Information (NI), neither self label nor counterpart's label disclosed; Private Information (PrI), only self label disclosed; Public Information (PuI), only counterparts' label disclosed; and Full Information (FI), both labels disclosed.

The stag-hunt game.

We use a standard two–by–two stag-hunt in which each player simultaneously chooses between A (cooperate) and B (do not cooperate). The payoff structure captures a classic coordination problem with two pure-strategy equilibria: a Pareto-superior but risky outcome (A,A) and a safer, risk-dominant but inefficient outcome (B,B), alongside two asymmetric non-coordination outcomes (A,B) and (B,A) (see Van Huyck et al. (2018)).

The stag-hunt is considered a trust-sensitive environment because achieving the socially efficient outcome (A,A) requires players to trust their counterpart will also choose A. If a player doubts their counterpart's cooperation, the rational best response is to select the safer option B, making cooperation contingent on mutual expectations and trust (Charness, 2000; Skyrms, 2003; Zhou et al., 2018). In repeated stag-hunt games, players continuously update beliefs based on observed behaviour, further highlighting the importance of trust and expectation formation (Thomas et al., 2014; Van Huyck et al., 1990).

This paradigm is particularly well suited to our research questions for two reasons. First, it allows us to conceptually and empirically distinguish between coordination—both players choosing the same strategy—and cooperation, both players choosing the socially efficient but risky option A. Second, since decisions critically depend on beliefs about the counterpart's choice, the stag-hunt provides a natural setting to examine how trait visibility impacts expectations and strategic behaviour.

Task 6. *Ex post* questionnaire (non-incentivised).

The final task in the experiment elicits *ex post* beliefs about participants' own personality type, first-order beliefs with respect to a mistrusting type, first-order beliefs with respect to a trusting type, second-order beliefs with respect to a mistrusting type, and second-order beliefs with respect to a trusting type. This task is used for informative purposes in the analyses.

Experimental procedures

The experiment was conducted at the Laboratory of Experimental Economics of Paris (LEEP) between March and April of 2023. We ran eight sessions with 192 voluntary participants (52.6% female) with an average age of 32 years (SD 13), consisting of mostly students (52%) and employed individuals (33%). Detailed information about the scholarity and occupations of participants are reported in the Appendix Tables 1.9 and 1.10. Average earnings were 16.16 euros (SD 3.17), including a 5 euro show up fee.

Across all rounds and treatments, the dataset contains 9,408 player–round observations (2,304 per treatment). Because these repeated observations are nested within participants, they are not statistically independent. Accordingly, in all regressions the unit of observation is the

player–round, and standard errors are clustered at the participant level to account for within-subject correlation.

1.4 Descriptive statistics

1.4.1 Beliefs and cooperation before information

We first describe behaviour in the one–shot baseline round (Task 2), collected before any mention of labels or visibility of information. The rate of individual cooperative choices (A in the stag-hunt) are high and statistically indistinguishable by self type (TT = 0.792; MT = 0.823; diff. = 3.1 pp; $p = 0.583$). Two–sample t–tests show no significant gap in second–order beliefs ($t = 1.43$, $p = 0.156$) and only a modest difference in first–order beliefs ($t = 2.07$, $p = 0.040$), with mistrusting types (MT) expecting slightly higher cooperation than trusting types (TT). This small difference, however, goes in the opposite direction of theoretical expectations, and disappears in a regression analysis, where *ex ante* type show no effects ($\beta = 0.005$, $p = 0.321$), nor it predicts FOB ($p = 0.583$) or SOB ($p = 0.993$). These results indicate that, before trait information is made salient, personality labels are neither behaviourally nor cognitively discriminating, as both trusting and mistrusting participants behave similarly and hold comparable expectations about others (see Table 1.2).

Table 1.2: Baseline round results by beliefs by type

	n	Cooperation	FOB	SOB
Mistrusting types	96	0.823	65.156	59.844
Trusting types	96	0.792	56.875	54.531
Diff. (MT−TT)		0.031	8.281	5.312
p-value		0.583	0.040	0.156

We next summarise *ex ante* beliefs from the elicitation questionnaire (Task 4), registered before informing participants about the experimental labels. Figure 1.3 plots the distribution of participants' *ex-ante* beliefs about their personality type, meaning the self–assessed probability (0–100) of being a TT vs. MT. Distributions align with the inventory classification: those later classified as MT report probabilities concentrated near 0, whereas TT concentrate near 100; kernel densities are single–peaked in the expected direction. Overall, participants' predictions on their personality types aligned with the inventory–based categorisation (Task 2).

Figure 1.3: *Ex ante* beliefs about personality type

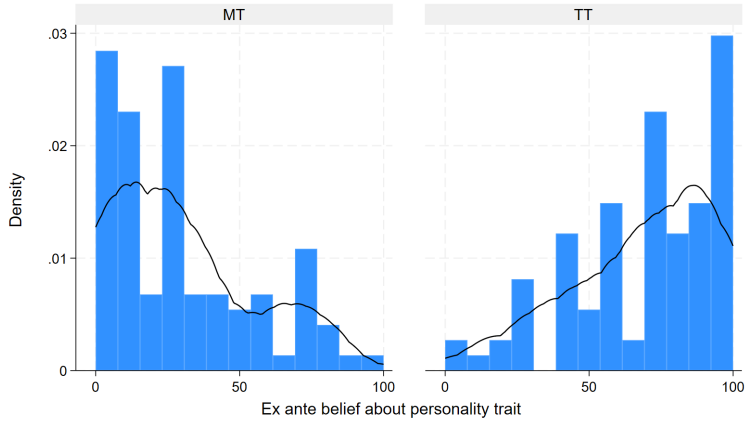


Table 1.3 summarises *ex ante* first–order and second–order beliefs both by self type and by the evaluated counterpart type. Two observations stand out: before any information about labels is introduced, participants expect somewhat more cooperation from TT than from MT counterparts (FOB ≈ 54 vs. 50; SOB ≈ 57 vs. 48). Among the counterpart–specific comparisons, the only difference reaching conventional significance is SOB when evaluating MT counterparts ($p = 0.039$). Second, overall FOB and SOB are slightly higher for MT than TT types (FOB diff. = 4.87, $p = 0.028$; SOB diff. = 5.81, $p = 0.021$), but these are modest in magnitude. In short, priors tilt towards expecting more cooperation from TT counterparts, while respondent–type differences are small at this stage.

Table 1.3: *Ex-ante* beliefs by self and counterpart type

	ex-ante FOB			*ex-ante* SOB		
Self type	Other: TT	Other: MT	Overall	Other: TT	Other: MT	Overall
MT	56.15 (27.62)	52.97 (26.43)	54.56 (13.09)	58.85 (26.61)	51.61 (25.07)	55.23 (14.16)
TT	52.66 (30.24)	46.72 (27.01)	49.69 (17.03)	55.10 (30.02)	43.75 (27.16)	49.43 (19.90)
MT − TT	3.49	6.25	4.87	3.75	7.86	5.81
p-value	0.405	0.107	0.028	0.361	0.039	0.021
n	192	192	192	192	192	192

1.4.2 Beliefs and cooperation after information

Figure 1.4 reports the mean rates of cooperative choices in the repeated game for each information regime. The highest value is in No information (0.79) and drops progressively with greater visibility to Public information (PuI) (0.684), Private information (PrI) (0.618) and Full information (FI) (0.554). Pairwise comparisons using two-sample t-tests confirm that growing trait visibility lowers cooperation relative to the baseline (NI), all significant at $p < 0.001$, with cooperation declining by -17% in PrI compared to NI (t = 12.98, $p < 0.001$) and by -24% in FI compared to NI (t = -17.63, $p < 0.001$).

Figure 1.4: Average decisions to cooperate by information regime

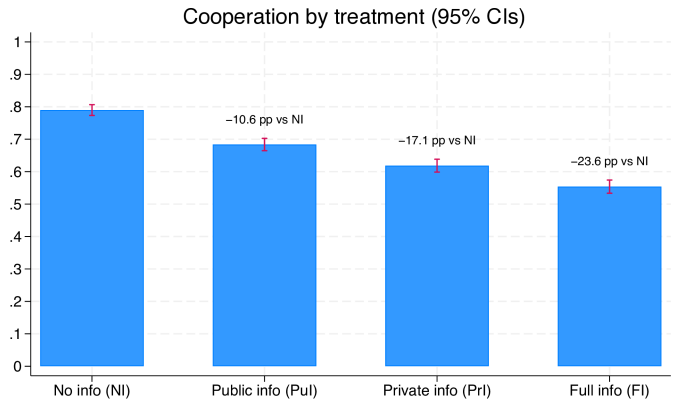
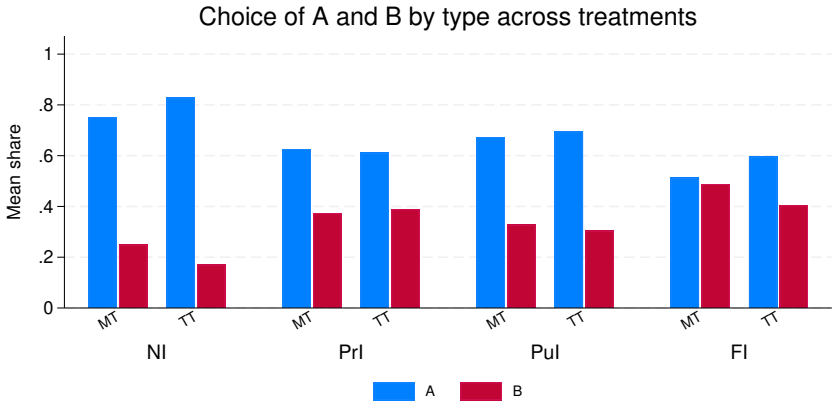


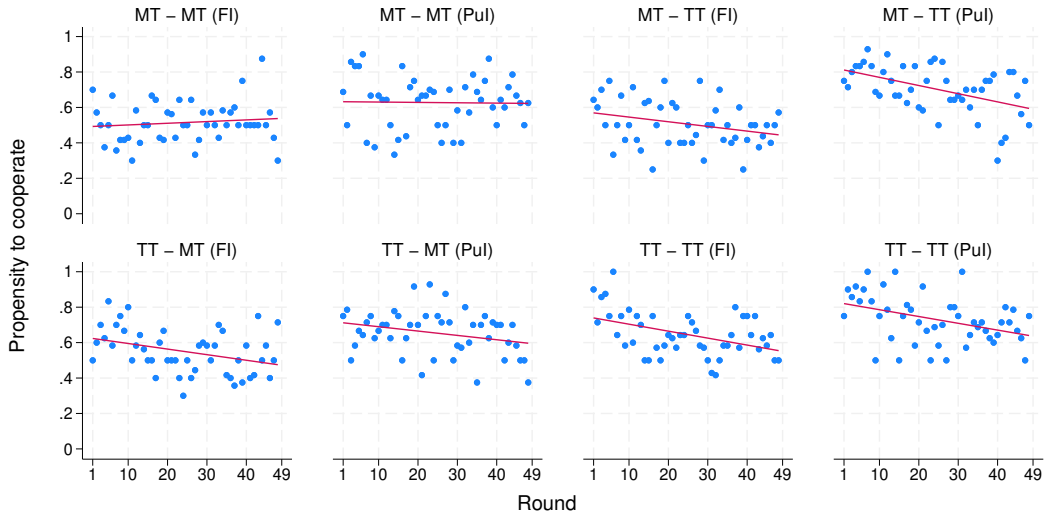
Figure 1.5 reports, within each information condition, the mean shares of cooperative (A) and non-cooperative choices (B) by participant type. The graph suggests that increasing type visibility reduces the willingness to choose A, i.e. the risky efficient action. Moreover, within any given condition, differences in choosing A or B are marginal between TT and MT types, but they do change between information regimes, suggesting that visibility was the dominant source of variation in average A choices at this level of aggregation.

Figure 1.5: A and B choices in the stag-hunt by type and treatment



When focusing on cooperation rates in the regimes where counterpart labels were visible (PuI + FI), Figure 1.6 summarises cooperative choices across the 48 rounds of the repeated game by self-type when paired with TT vs MT counterparts. The graph further shows the marginality in differences across types and that the main variation occurred across treatments. The graph shows that type-based stereotype did not produce significant differences in actual cooperation between TT and MT (all declining overtime except for MT-MT pairings), but rather label visibility lowered cooperation across the board.

Figure 1.6: Propensity to cooperate when counterpart label visible



When turning to first-order beliefs (FOB), Figure 1.7 shows that revealing the counterpart's personality label does not substantially reshaped participants' expectations, as first-order beliefs remain remarkably similar across visibility conditions and change only modestly in the predicted direction (slightly higher expected cooperation from TT than MT). Again, suggesting that type-based discrimination is not the primary driver of the drop in cooperative choices. Overall, the graph suggests a wide decline in anticipating cooperation rather than the type specific rise one would normally anticipate from TT.

Figure 1.7: FOB by type when counterpart label is visible

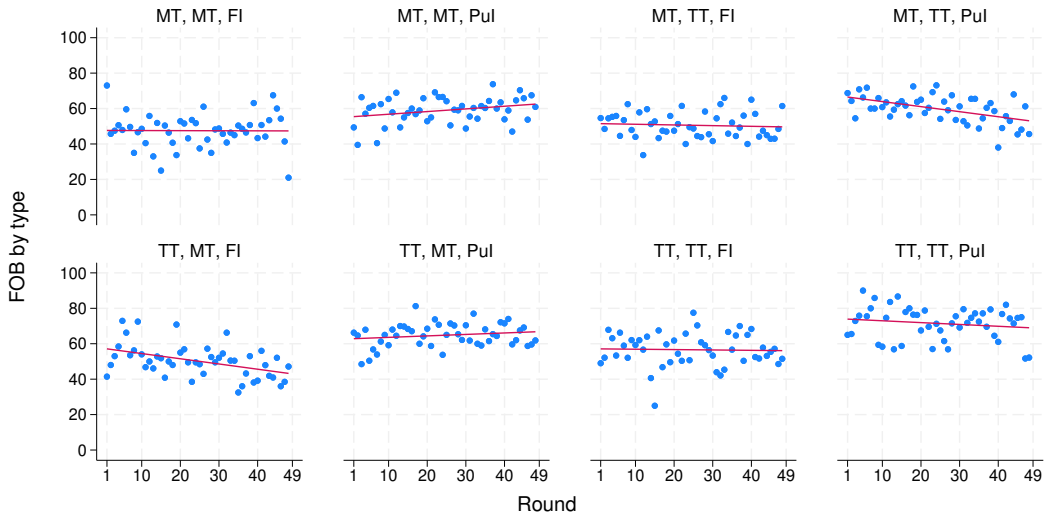
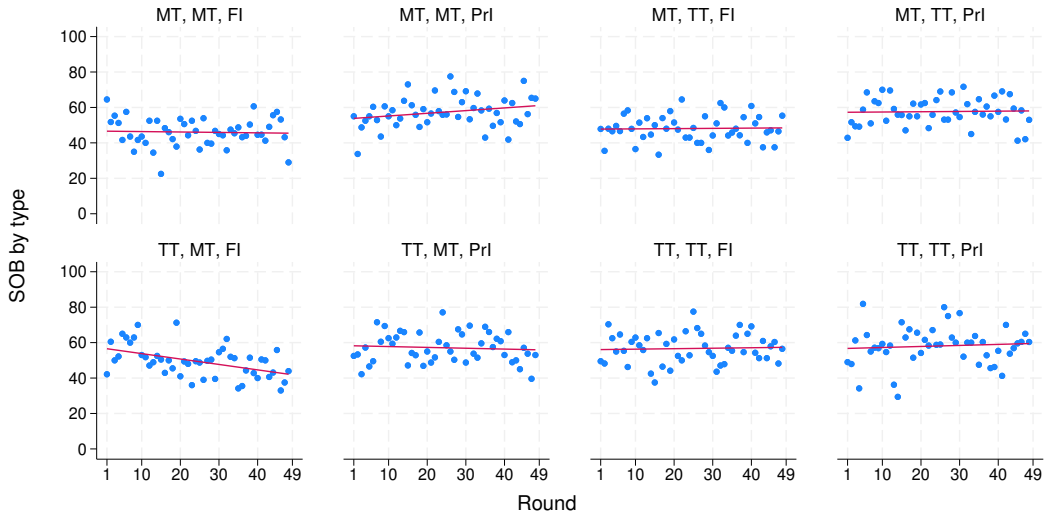


Figure 1.8 plots participants' second-order beliefs (SOB) across regimes in which the participant's own label was visible (PrI + FI). Overall, SOB shows only modest variation across visibility conditions: SOB remain broadly stable and any differences between TT and MT are

small. This suggests that disclosing one's own label does not dramatically change what individuals think others expect of them. Importantly, the modest movement in SOB contrasts with the large drop in cooperation in the relevant information regimes, indicating that participants do not substantially revise their SOB, yet they become less willing to choose the cooperative action. This suggests that self-label visibility alters strategic incentives and equilibrium selection without large shifts in what people think others expect of them.

Figure 1.8: SOB by type when self label is visible



1.4.3 Coordination across information regimes

We examine overall strategic coordination across information regimes, defined as the share of rounds in which both players choose the same action (either A or B). This captures alignment in strategy choice irrespective of whether their choices settles on the efficient or the inefficient equilibrium. Table 1.4 reports average coordination rates: the bottom row presents the Excess Coordination (EC) index, which measures the extent to which observed coordination exceeds the level expected under independent play (Expected Coordination, row 3). While raw coordination rates are relatively high across information regimes, the EC index reveals substantial differences in the degree of strategic alignment.

To interpret these levels, we benchmark them against expected coordination under independent play. Let s_A denote the observed cooperation rate (row 2, i.e., the share of A choices). If players chose independently given s_A, the chance they match on the same action is $s_A^2+(1-s_A)^2$ (row 3). We then report the EC index that scales how much observed coordination exceeds this benchmark: $\text{EC} = \dfrac{\pi_{AA}+\pi_{BB}-\left(s_A^2+(1-s_A)^2\right)}{1-\left(s_A^2+(1-s_A)^2\right)}$, where $\pi_{AA}+\pi_{BB}$ is the observed coordination rate (row 1). EC equals 0 under independence and approaches 1 as coordination nears its upper bound. For FI, $s_A = 0.554$ implies expected coordination 0.506; with observed coordination 0.757, $\text{EC} = (0.757-0.506)/(1-0.506) = 0.508$ (row 4).

While raw coordination is high in all treatments, EC reveals sizeable differences in strategic alignment. FI has by far the strongest strategic alignment, even though it has the lowest

cooperation. This indicates that label visibility helps players converge on a safe strategy. On the opposite end, in NI the natural focal point is A, the payoff-dominant equilibrium. These results are consistent with previous experimental findings showing that increased information enhances mutual predictability and facilitates convergence toward coordinated outcomes (see Abeler et al. (2019); Devetag and Ortmann (2007); Giorgetta et al. (2021).

Although raw coordination is high across all information regimes, the EC index reveals meaningful differences in strategic alignment. FI exhibits the strongest excess coordination, despite having the lowest cooperation rate, indicating that label visibility helps players converge on a safe, predictable strategy. These findings are consistent with prior experimental evidence showing that increased information improves mutual predictability and facilitates convergence toward coordinated outcomes (Abeler et al., 2019; Devetag and Ortmann, 2007; Giorgetta et al., 2021).

Table 1.4: Coordination and cooperation across information regimes

	NI	PrI	PuI	FI
Coordination	0.717	0.704	0.647	0.757
Cooperation	0.790	0.618	0.684	0.554
Expected Coordination	0.668	0.528	0.568	0.506
Excess Coordination (EC)	0.147	0.373	0.183	0.508

1.5 Results: How information reshapes beliefs and cooperation

1.5.1 Testing Mechanism 1: Self-identification

To test Mechanism 1, we examine whether behaviour is influenced by the activation and possible adjustment of beliefs about one's own personality type. The mechanism posits that individuals internalise the type they associate with, especially when it becomes salient in the experimental context.

In our design, self-identification unfolds in two stages. First, prior to any label revelation, participants predict whether they belong to the "trusting" or "mistrusting" type. This task makes the personality dimension salient and may already affect behaviour via self-perception. Second, in regimes where the self-label is revealed (Private & Full information), participants additionally learn their inventory-based measured type, which may confirm or contradict the prior and trigger a revised self-identification. We model this process using these variables:

- **Believed type**: `exante_belief_type`, the participant's predicted probability (0-1) of being a mistrusting type. This captures prior self-perception before any label is revealed (Task 4, question 1).
- **Informational shock**: `info_type`, the measured type minus `exante_belief_type`. It measures how much the information diverges from expectations: Positive = more MT

than expected; negative = more TT than expected.

- **Visibility of self-label**: `self_label_visible`, information regimes where self-label is visible (Private information + Full information).
- **Visibility of counterpart label**: `other_label_visible`, information regimes where counterpart label is visible (Public information + Full information).
- **Type confirmation**: `info_type_self`, self-relevant informational shock when self label is visible. This captures whether the participant received new self-relevant information and, if so, in what direction (positive = more mistrusting than expected, negative = less). Additionally, we include `info_type_other` to account for `other_label_visible`.
- **FOB**: First-order beliefs, registered on a 0-100 scale.

We estimate logistic regressions of choosing A (cooperative choices, the dependent variable) in the stag-hunt including these variables plus demographic controls. A positive coefficient on `exante_belief_type` would indicate that identifying as a trusting type based on one's own belief increases the likelihood of cooperation. A significant coefficient on the visibility-weighted shock terms (`info_type_self`, `info_type_other`) would indicate that receiving self-relevant or regime-relevant information about the measured type changes behaviour in the direction of the informational shock.

Table 1.5 summarises the main findings. Column 1 shows the results from regressing cooperative choices in the one-shot baseline round before any information is introduced (one choice per participant $N = 192$): neither `exante_belief_type` nor `info_type` predict cooperation, however, *ex-ante* FOB is positively associated with cooperative choices (0.043, $p < 0.010$). In column 2, we re-estimate the model using only the first round of the repeated game, immediately after labels are introduced and before any feedback (only in regimes where self-label is visible PrI and FI; $N = 96$). Results show that only FOB significantly predict cooperation (0.026, $p < 0.010$).

Column 3 reports results for the repeated stag-hunt with participant-clustered standard errors and round fixed effects. FOB remains a strong and stable predictor (0.051, $p < 0.010$), whereas `exante_belief_type`, `info_type`, and the visibility-weighted shock variables are not significant. Similarly, the label visibility variables are not statistically distinguishable from zero: self-type visible (0.059, $p = 0.906$); other-type visible (-0.348, $p = 0.455$), and their interaction (0.322, $p = 0.511$).

In sum, results from Mechanism 1 show that while disclosing self labels shape players' expectations about others, therefore driving their own cooperative choices, there is no direct evidence that label visibility entices the internalisation of the type associated with said label, thereby rejecting Mechanism 1.

Table 1.5: Mechanism 1: Self-identification

Variables	Baseline stag hunt	Round 1	Repeated stag hunt
FOB	0.043***	0.026**	0.051***
	(0.009)	(0.011)	(0.004)
exante_belief_type	-0.289	0.365	0.301
	(0.570)	(0.598)	(0.340)
info_type	-0.304	0.370	0.304
	(0.576)	(0.602)	(0.344)
info_type_self	0.004	–	0.008
	(0.014)		(0.007)
info_type_other	–	–	-0.003
			(0.008)
self_label_visible	-0.392	–	0.069
	(0.866)		(0.496)
other_label_visible	0.382	–	-0.348
	(0.708)		(0.461)
Self × Other visible	–	–	0.325
			(0.894)
Age	-0.000	-0.002	-0.007
	(0.016)	(0.018)	(0.009)
Gender (Female = 1)	0.566	-0.326	0.385
	(0.435)	(0.480)	(0.247)
Occupation	-0.003	-0.164	0.007
	(0.130)	(0.138)	(0.081)
Risk aversion	-0.013	0.112	-0.023
	(0.095)	(0.110)	(0.068)
Scholarity	0.129	0.412*	-0.006
	(0.171)	(0.216)	(0.099)
Observations	192	96	9,216

Standard errors in parentheses. ***, **, * indicate $p < 0.01$, $p < 0.05$, $p < 0.10$. Repeated game: participant-clustered SEs with round fixed effects. For Baseline and Round 1 only robust SEs.

1.5.2 Testing Mechanism 2: Type-based discrimination.

We now turn to Mechanism 2, which posits that when the counterpart's personality label is visible, participants may condition cooperation on the perceived desirability of that type. This mechanism captures two possible behavioural biases: (1) a preference for coordinating with a preferred type (e.g., trusting individuals), and (2) a tendency to favour counterparts of the same type, a form of homophily. Both effects emerge only when the counterpart's type is visible, so we use the following variables:

- **Self TT**: `self_type`, equals 1 if the participant's self measured trait in the personality questionnaire (Task 3) is Trusting (TT), 0 if Mistrusting (MT).
- **Counterpart TT**: `type_other`, equals 1 if the counterpart's measured trait in the personality questionnaire (Task 3) is Trusting (TT), 0 if Mistrusting (MT).

We estimate two logistic regression models of choosing A, reported in Table 1.6. In Model Counterpart type preference (Column 1), we delimit observations to information regimes where counterpart label is visible (FI + PuI, N=4,608), including `type_other`, `self_label_visible` and their interaction to check whether the effect of counterpart type differs when self-label is

not visible (PuI vs FI). Results show that FOB is strongly predictive of cooperation (0.051, $p < 0.001$), however, the main effect of facing a trusting counterpart is positive but not statistically significant (0.257, $p = 0.205$), same as the self-visibility variable (0.172, $p = 0.628$), while the interaction is small and not significant (-0.232, $p = 0.394$).

In Model Homophily (Column 2), we include `self_type` and restrict observations to the Full information regimes, as both types must be visible to test whether preferences for similar others exist. Results show that while FOB remains strongly predictive (0.058, $p < 0.001$), neither `self_type` nor `type_other` have a significant effect ($= 0.069$, $p = 0.901$; $= -0.235$, $p = 0.379$), while the homophily interaction is positive but not statistically significant ($= 0.518$, $p = 0.179$).

Across both models, results do not provide statistically significant evidence of type-based discrimination or homophily in cooperation, instead FOB remain the dominant and stable predictor of cooperative choices when counterpart labels are visible.

Table 1.6: Mechanism 2: Type-based discrimination

Variables	**Counterpart type preference**	**Homophily**
`type_other` is TT	0.257	−0.235
	(0.203)	(0.267)
FOB	0.051***	0.058***
	(0.005)	(0.009)
`self_label_visible`	0.172	–
	(0.355)	
Counterpart TT×self	−0.232	–
	(0.273)	
`self_type`	–	0.069
		(0.554)
`self_type`×**Counterpart TT**	–	0.518
		(0.386)
Gender (Female = 1)	0.077	−0.620
	(0.352)	(0.503)
Age	0.013	0.019
	(0.013)	(0.020)
Occupation	0.062	0.131
	(0.114)	(0.167)
Scholarity	0.059	0.062
	(0.132)	(0.189)
Risk aversion	0.081	0.142
	(0.079)	(0.139)
Constant	−3.807**	−3.723**
	(1.330)	(1.830)
Observations	4,608	2,304

Logit; participant–clustered SEs in parentheses; ***, **, * $p < 0.01, 0.05, 0.10$.

1.5.3 Testing Mechanism 3: First-order belief bias.

We now test Mechanism 3, which proposes that personality labels not only guide strategic behaviour, but also distort expectations. Specifically, when the counterpart's type is visible, participants may hold overly optimistic beliefs about the likelihood of cooperation from those labelled as trusting types (TT), independent of actual behaviour. This belief-based distortion

can itself become self-fulfilling, as previously shown: higher first-order beliefs (FOB) significantly increase the probability of cooperation. We estimate three OLS models with FOB as the dependent variable, restricting samples to rounds in which the counterpart's label is visible and use the following regressors:

- **counterpart_cooperates**: Whether the counterpart chose A in the stag-hunt.
- **Counterpart TT** (`type_other`): = 1 if the counterpart's measured trait is Trusting (TT), 0 if Mistrusting (MT).
- **Self-label visibility** (self_label_visible): = 1 if the participant's own label is visible (Private or Full), 0 otherwise.

We first report results from regressing FOB on the three regressors in the repeated stag-hunt game (see Table 1.7). Results show that a cooperating counterpart strongly raises beliefs (+25.30, $p < 0.001$); conditional on behaviour and visibility, knowing a counterpart's label is non-significant ($p = 0.374$; `type_other` × `self_label_visible`: $p = 0.902$), while self-label visibility is associated with lower beliefs (−9.08, $p = 0.029$).

We then isolate a purely label–driven association by restricting the analysis to round 1 of the repeated stag-hunt, when no feedback has yet been observed. We regress FOB on `type_other`, `self_label_visible` and their interaction. Results show a positive, marginal association for `type_other` (+15.60, $p = 0.069$) and a negative, marginal interaction (−20.35, $p = 0.085$), consistent with initial optimism toward TT counterparts that attenuates when the self label is also visible. This result show a small, marginal bias in favour of TT counterparts, which fades when self label is also visible.

To address simultaneity (beliefs and behaviour moving together within a round), we regress FOB on the counterpart's cooperation in the previous round (`partner_cooperates_lagged`). The objective is to test whether beliefs track recent experience in the repeated game once initial label effects fade. In this estimation (Column 3), the lagged outcome remains significant (+20.42, $p < 0.001$), while `type_other` and `type_other` × `self_label_visible` are not. Overall, this estimation shows that labels exert a short-lived effect on beliefs at the very start, however, participants update expectations based on what their counterpart actually did in the previous round.

In sum, results weakly confirm Mechanism 3, as label visibility initially inflate first-order beliefs in favour of trusting types, which increases cooperation with them and confirm the existence of bias. However, this FOB-based bias fades with feedback, after which beliefs primarily reflect observed behaviour.

Table 1.7: Mechanism 3: First-order belief bias

Variables	Repeated stag hunt	Round 1	Lagged behaviour
counterpart_cooperates	25.30***	–	–
	(2.720)		
type_other	3.053	15.600*	2.680
	(3.417)	(8.465)	(3.771)
self_label_visible	-9.084**	-0.971	-8.325*
	(4.107)	(8.790)	(4.274)
type_other × self_label_visible	-0.526	-20.35*	2.044
	(4.262)	(11.690)	(4.849)
partner_cooperates_lagged	–	–	20.420***
			(2.579)
Age	-0.085	-0.618**	-0.041
	(0.133)	(0.252)	(0.137)
Gender (Female = 1)	2.309	-7.087	4.297
	(3.625)	(6.006)	(3.711)
Occupation	-2.472**	0.355	-2.688**
	(1.057)	(2.060)	(1.043)
Risk aversion	-0.427	0.293	-0.831
	(0.994)	(1.465)	(0.932)
Scholarity	1.555	1.431	2.037
	(1.723)	(2.102)	(1.599)
Constant	46.07***	77.16***	44.31***
	(14.35)	(23.47)	(13.92)
Observations	4,608	96	4,512
R-squared	0.198	0.164	0.255

Robust standard errors in parentheses. ***, **, * indicate $p < 0.01, , p < 0.05, p < 0.10$. Column (3) absorbs round and match/group fixed effects and clusters by participant.

1.5.4 Testing Mechanism 4: Second-order belief pessimism.

We now examine Mechanism 4, which posits that participants adjust their expectations about how cooperative others expect them to be, and that this meta-belief influences their own willingness to cooperate. Intuitively, if a mistrusting participant (MT) believes their counterpart knows their type and therefore expects them to defect, the participant may lower their own willingness to coordinate.

We estimate a linear regression with the participant's second-order belief (SOB), that is, their belief about how likely the counterpart thinks they will cooperate, as dependent variable. We restrict the analysis to regimes where counterpart's label is visible (`other_label_visible`), and include the counterpart's FOB (`fob_counterpart`) to isolate pessimism specifically related to how participants believe they are perceived beyond any general variation in beliefs. As with Mechanism 3, repeated feedback allows them to form expectations over how they are likely perceived.

The variables of interest mirror those used in the self-identification test (Mechanism 1): `exante_belief_type`, `info_type` and `info_type_self`. In this context, these variables capture the extent to which participants anticipate being perceived through the lens of their measured type, regardless of whether they personally endorse it. This means that they reflect expectations about how their personality label—once made visible—will be interpreted by the counterpart, especially when the it is associated with mistrusting traits.

The results in Table 1.8 shows that `fob_counterpart` is a strong predictor of SOB (0.315, $p < 0.001$), indicating that meta-beliefs move with how optimistic the counterpart is about the participant's cooperation. By contrast, `exante_belief_type` and `info_type` are not statistically significant (both $p > 0.180$). The visibility-weighted surprise term `info_type_self` is small but positive (0.219, $p = 0.044$), suggesting a modest adjustment of SOB when feedback about one's type diverges from priors and is personally visible. The main effect of `self_label_visible` is not significant.

In sum, results show limited support for systematic second–order belief pessimism driven directly by labels, as effects seem contingent on self–visibility. Combined with earlier results showing that FOB is the proximate driver of cooperation, this suggests that reputational concerns matter for when accounting for the beliefs held by the counterparts, rather than via a strong, direct pessimism channel in SOB.

Table 1.8: Mechanism 4: Second-order belief pessimism

Variables	**SOB**
`fob_counterpart`	0.315***
	(0.041)
`exante_belief_type`	5.665
	(4.243)
`info_type`	5.602
	(4.286)
`info_type_self`	0.219**
	(0.107)
`self_label_visible`	2.367
	(6.212)
Age	-0.215*
	(0.124)
Gender (Female=1)	2.891
	(3.425)
Occupation	-2.687**
	(1.003)
Risk aversion	-0.653
	(0.951)
Scholarity	2.065
	(1.609)
Constant	39.79***
	(13.94)
Observations	4,608
R^2	0.213

Robust standard errors in parentheses; $p < 0.01, , p < 0.05, p < 0.10$.

1.6 Concluding Remarks

In this paper, we set out to study how the visibility of personality traits influences coordination and cooperation in a repeated stag-hunt game. We focused on four psychological and strategic mechanisms through which information may shape behaviour: self-identification, type-based discrimination, first-order belief bias, and second-order belief pessimism. By eliciting behavioural

choices and beliefs round by round, we identified the causal impact of different forms of trait visibility on expectations and outcomes.

The findings show that personality labels primarily affect expectations rather than preferences. In the absence of information, personality traits are not predictive of behaviour: trusting and mistrusting types coordinate at similarly high rates. However, when labels are visible, players initially favour trusting types, generating assortative cooperation. This effect, nonetheless, fades as the game progresses, suggesting that type-based discrimination does not produce sustained behavioural conformity.

Overall, our findings show that beliefs are the proximate drivers of choices. They also highlight a "two-edged" effect from revealing personality information, reconciling prior mixed evidence on the matter: visibility does not uniformly promote or hinder collective efficiency, but it simultaneously enhances predictability while undermining trust in others. This is illustrated in how cooperation declines sharply and progressively as label visibility increases as well as in how full visibility captures the steepest fall in cooperative choices and the highest coordination around safe choices at the same. Recognising this duality is essential for settings where personality cues or identity markers are made salient, particularly within organisational behaviour.

In sum, our findings imply that interventions aiming to improve group outcomes should prioritise behaviour-linked feedback and reputation signals over permanent trait disclosure when the policy objective is sustained cooperation.

Limitations and future directions. While our analyses isolate four distinct mechanisms, we do not assume they are fully orthogonal. In practice, these mechanisms may interact or reinforce each other. A unified empirical strategy would help clarify the causal sequencing and interdependence of beliefs and choices. However, combining decisions and beliefs in the same framework poses technical and conceptual challenges, especially in repeated settings.

Second, we recognise certain tweaks that could have enhanced our experimental setting to increase significance in the results. First, participant categorisation might have misclassified certain players and attenuated true type effects in beliefs and behaviours. A possible way to tackle this effect in future research could be screening participants *ex-ante* or do a tighter categorisation by keeping those with scores around the edges only. Second, revealing information could make the stag-hunt game riskier, as players could expect others to condition their choices on labels, thereby making mutual assurance becomes harder. This could have made choosing *A* feel more exposed than usual in this setting, while making non-cooperation overly attractive. Similarly, changing the payoff matrix or adapting the experimental game could be potential avenues to explore.

Contribution and originality. While mechanisms such as self-identification and type-based discrimination have been well documented in previous research on social identity and economic behaviour, our study is among the first to empirically isolate and test the role of belief-based distortions arising from trait visibility. Specifically, we show that participants form biased first-order beliefs about counterparts based on their personality labels, even when controlling for actual behaviour. This first-order belief bias (Mechanism 3) reflects a form of stereotype-driven expectation not previously captured in coordination games.

This layered belief structure allows us to reframe coordination as an epistemic process: who

sees what, and with what visibility, determines which beliefs are formed and which equilibria are selected. Personality trait visibility operates as a double-edged signal: it enhances predictability, but triggers belief distortions and social inferences that degrade trust. In strategic environments increasingly shaped by observable traits-be it personality, group markers, or else-understanding the belief-mediated effects of identity activation is essential.

To broaden the relevance of our results, future work should examine how these mechanisms operate outside formal game settings. First, the first-order belief bias (Mechanism 3) can be read as a stereotype-based predictive bias: people form expectations about others from visible labels rather than from experience or interaction. Second, the second-order belief pessimism (Mechanism 4) under mutual visibility maps onto anticipated misrecognition: people lower their own expectations and adjust behaviour because they expect to be judged through a social category.

Furthermore, changing the stakes and payoff geometry might advance our lines of research, by testing whether higher stakes or different coordination games (e.g. assurance games) alter the predictability/cooperation trade-off. If revealing labels pushed toward safety, larger payoffs might amplify or attenuate that tendency.

These generalised formulations open the door to interdisciplinary exploration. They suggest that the effects we observe in trait-labelled stag-hunt games may also operate in domains such as employment decisions, classroom dynamics, or political discourse—whenever individuals are classified into visible categories with presumed behavioural traits.

1.7 Appendix

Demographic data

Table 1.9: Demographics on scholarity

Diplomas	session 1	session 2	session 3	session 4	session 5	session 6	session 7	session 8
PhD (8 years)	1	0	0	1	0	0	0	0
Master (5 years)	5	7	4	8	4	7	2	6
Bachelor (3 years)	8	9	4	8	11	6	4	7
Technical degree (2 years)	4	4	4	1	3	5	7	5
High school	6	3	11	5	6	5	11	4
Secondary education	0	0	1	1	0	1	0	2
No diploma	0	1	0	0	0	0	0	0

Table 1.10: Demographics on occupation

Occupation	session 1	session 2	session 3	session 4	session 5	session 6	session 7	session 8
Student	14	10	13	11	18	10	14	10
Employed	8	11	7	6	5	7	7	12
Unemployed	2	1	0	1	1	4	1	1
Pensioner	0	0	4	1	0	1	1	0
Other	0	2	0	5	0	2	1	1

Descriptive Statistics

Table 1.11: *Ex ante* vs *ex post* beliefs on coordination

		FOB on a TT	FOB on a MT	SOB on a TT	SOB on a MT
Ex ante	Overall	54.401	49.843	56.98	47.682
	Trusting type	52.656	46.718	55.104	43.75
	Mistrusting type	56.145	52.968	58.854	51.614
Ex post	Overall	60.572	50.781	59.322	52.708
	Trusting type	62.604	47.239	62.031	53.645
	Mistrusting type	58.541	54.322	56.614	51.770

Hybrid personality questionnaire

We present the hybrid use of the questionnaire used in the experiment, drawn from the NEO-PI (facet: Trust) and HEXACO (facets: Forgiveness, Gentleness, Flexibility). For classification, the composite score (mean across all items) was dichotomised using the session median: participants with a composite score above the session median are classified as *trusting personalities*, those below as *mistrusting personalities*.

Instructions

For each statement below, indicate how much you agree or disagree using the following scale: 1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree (neutral), 4 = Agree, 5 = Strongly agree.

Table 1.12: Personality questionnaire items

	# Item	Source (Inventory / Facet / original item #)
1	I believe that most people are fundamentally well-intentioned.	NEO / Trust / 34
2	My attitude toward people who have treated me unfairly is to "forgive and forget."	HEXACO / Forgiveness / 27
3	When someone tells me I am wrong, my first reaction is to defend my point of view. (R)	HEXACO / Flexibility / 63 (R)
4	I tend to assume the best about other people.	NEO / Trust / 184
5	I believe that most people will take advantage of me if I let them. (R)	NEO / Trust / 64 (R)
6	I generally accept other people's faults without complaining about them.	HEXACO / Gentleness / 33
7	If someone has cheated me before, I will always be suspicious of them. (R)	HEXACO / Forgiveness / 51 (R)
8	I am fairly flexible in my opinions when people disagree with me.	HEXACO / Flexibility / 39
9	I tend to be cynical and skeptical about other people. (R)	NEO / Trust / 4 (R)
10	I am rarely resentful, even toward people who have caused me serious harm.	HEXACO / Forgiveness / 3
11	I find it difficult to compromise when I truly believe I am right. (R)	HEXACO / Flexibility / 87 (R)
12	I am suspicious when someone does something kind for me. (R)	NEO / Trust / 124 (R)
13	My first reaction is to trust people.	NEO / Trust / 154
14	Sometimes people tell me I judge others too harshly. (R)	HEXACO / Gentleness / 9 (R)
15	Even when people make many mistakes, I rarely make negative comments about them.	HEXACO / Gentleness / 81
16	I think most people to whom I am connected are honest and reliable.	NEO / Trust / 94
17	I find it hard to completely forgive someone who has hurt me. (R)	HEXACO / Forgiveness / 75 (R)
18	I often judge others with leniency.	HEXACO / Gentleness / 57
19	People sometimes say I am stubborn. (R)	HEXACO / Flexibility / 15 (R)
20	I have great faith in human nature.	NEO / Trust / 214

Item-to-facet mapping (summary)

Agreeableness (HEXACO & NEO-PI): High scores indicate a tendency to forgive, make lenient judgments of others, accept compromises, and control temper; low scores relate to holding grudges, critical evaluations of others, stubborn defence of one's viewpoint, and readiness to feel anger after mistreatment.

Forgiveness (HEXACO): Assesses the inclination to forgive and to restore trust after being wronged. Low scorers are more likely to hold a grudge; high scorers more readily reinstate friendly relations.

- Items: #2, #7 (R), #10, #17 (R).

Gentleness (HEXACO): Assesses willingness to accommodate others and to compromise. Low scorers are more opinionated and argumentative.

- Items: #14 (R), #6, #18, #15.

Flexibility (HEXACO): Assesses belief in the goodwill of others and readiness to trust. High scorers expect honesty and reliability; low scorers tend to be cynical or suspicious.

- Items: #19 (R), #8, #3 (R), #11 (R).

Trust (NEO-PI-R): tendency to assume good intentions in others and to trust them.

- Items: #1, #4, #5 (R), #9 (R), #12 (R), #13, #16, #20.

Experimental Instructions

This subsection presents translated version of the experiment's instructions, from their original French presentation.

Page 1. Introductions to the experiment

The experiment consists of four parts:

1. A lottery choice followed by a single period of a game.
2. A personality questionnaire.
3. A series of predictions.
4. Forty-eight periods of a repeated game followed by a set of questions.

Please pay close attention to the instructions. The variable part of your payoff depends on your predictions, your decisions, and the decisions of other participants.

Part 1: One-shot game with lottery choice

In the first part, you will play a single period of a game with another participant randomly selected from the session. In this game, both you and your partner will make two predictions and one decision simultaneously. You will be asked to indicate:

- What you anticipate the other participant will choose (may be paid).
- What you believe the other participant anticipates about your decision (may be paid).
- Your own decision (paid).

Part 2: Personality questionnaire

In the second part, you will complete a questionnaire. Based on your responses, participants will be divided into two categories of 12 individuals each according to the similarity of responses. We will explain how the two categories are determined after the questionnaire. Please note that this part does not generate any earnings.

Part 3: Series of predictions

In the third part, you will make a series of predictions:

1. The category to which you belong.
2. Your prediction of other participants' decisions.
3. Your belief about what other participants anticipate regarding your decision.

Part 4: Repeated game

Finally, in the fourth part, you will play 48 periods of the game in a group of 6 participants, including 3 from each category. In each period, you will be randomly paired with another participant within your group. In each period of the repeated game, you will be asked to indicate:

1. Your prediction of the other participant's decision (may be paid).
2. Your belief about what the other participant anticipates regarding your decision (may be paid).
3. Your own decision (paid).

You will be informed of the outcome at the end of each period.

Payoff structure

Your final payoff will consist of:

- A fixed amount of €5.
- A variable amount depending on:
  - ECUs (experimental currency units) accumulated over the 49 periods of the game, based on your decisions and the decisions of others.
  - Euro earnings from your predictions of the other participant's decisions in one randomly selected period among the 49 periods.
  - Euro earnings from your predictions of the other participant's anticipation of your decisions in one randomly selected period among the 49 periods.
  - Earnings from the lottery choice.

Please note that you will make your lottery choice before the experiment begins, but the random draw that determines your lottery payoff will be conducted at the end of the experiment.

Page 2. Risk-attitude task

You will face 10 decisions displayed on your screen. Each decision is a choice between *Option A* and *Option B*. While the payoffs for each option are fixed across all decisions, the probability of obtaining a high payoff varies for each option.

After making all your choices, one of the 10 decisions will be randomly selected to determine your payment. For the option you chose in that decision (A or B), chance (according to the corresponding probabilities) will determine whether you receive the low or high payoff.

Summary: You will make 10 choices. For each decision, you must choose between Option A and Option B. You can choose A for some decisions and B for others.

To determine your earnings, one of the 10 decisions will be randomly selected for payment. Then, a number will be drawn to determine your payoff for the option you selected in that decision. Both random draws will be made at the end of the experiment.

Page 3. One-shot baseline round

You will play a single round of the game, which consists of making two predictions and one decision simultaneously with another randomly selected participant in the session.

You will indicate:

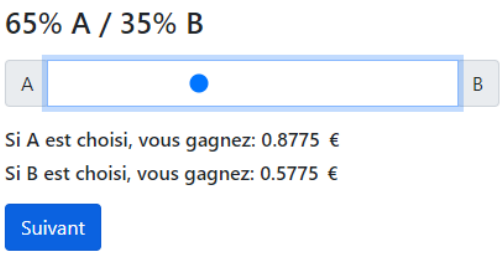
1. Your prediction of the other participant's decision.
2. Your prediction of the other participant's belief about your decision, i.e., what the other participant thinks you will choose.

For each prediction, you will use a slider to indicate the probability (in percentages) that the choice is A or B. By definition, the two probabilities sum to 100%. Moving the slider allows you to assign probabilities to both options simultaneously.

Below the slider, the payoffs corresponding to your predictions for the cases where the correct answer is A or B are displayed. The slider must be moved to proceed to the next page.

Earnings for decisions are expressed in *ECUs* and are accumulated throughout the experiment. Earnings for the two belief predictions will be paid in euros; two responses (one for each type of belief) will be randomly selected at the end of the experiment among all the beliefs you submitted.

To familiarise yourself with the interface, imagine that you need to predict the respective probabilities that the correct answer is A or B: move the slider to see the resulting payoffs for your prediction.



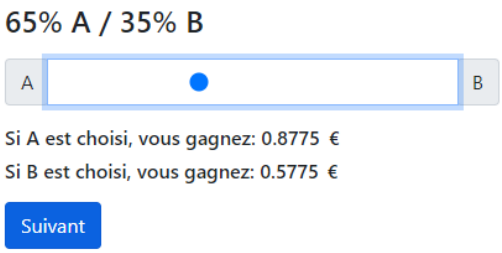
Page 4. Stag-hunt game

Next, you and the other participant must choose between decision *A* and decision *B*. The table below shows the ECUs payoffs corresponding to each combination of choices. Blue numbers indicate your payoff, black numbers indicate the other participant's payoff.

		L'autre participant	
		A	B
Vous	Je choisis A	6, 6	-3, 3
	Je choisis B	3, -3	3, 3

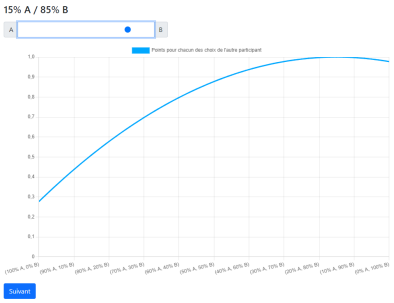
Page 5. FOB Baseline

You are paired randomly with another participant. Please indicate your probability estimate that the other participant will choose A or B in this round using the slider below. The euro payoffs for your prediction depending on the other participant's choice are displayed below the slider.



Page 6. SOB Baseline

The other participant has estimated the probability you assign to choosing A or B. Which response do you believe they gave?



Page 7. Stag-hunt Baseline

Please make your choice:

		L'autre participant	
		A	B
Vous	Je choisis A	6, 6	-3, 3
	Je choisis B	3, -3	3, 3

Page 8. Questionnaire

The personality questionnaire consists of 20 statements, each with a five-point response scale indicating your level of agreement.

- 1 = Strongly disagree
- 2 = Disagree
- 3 = Neutral
- 4 = Agree
- 5 = Strongly agree

There are no correct or incorrect answers. Read each statement carefully and indicate your true level of agreement by selecting a single response for each statement.

Page 9. Categorisation

Your responses will be used to compute a score to classify participants in this session as either *trusting* or *mistrusting*.

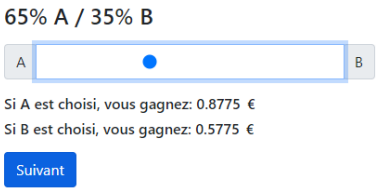
In this session, the 24 participants will be ranked by score: the 12 participants with the highest scores will be classified as *trusting*, the 12 with the lowest scores as *mistrusting*.

Mistrusting participants	Trusting participants
Sceptical about compromises and cooperation. They tend to believe in others' dishonesty, prefer competition to cooperation, are critical in social relations, and hold grudges when wronged.	Easily compromise and cooperate. They tend to believe in honesty and good intentions, are lenient in social interactions, and forgive easily if wronged.

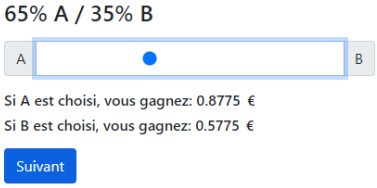
Pages 10-15. Ex-ante belief elicitation questionnaire

Over the following pages, you will make five predictions.

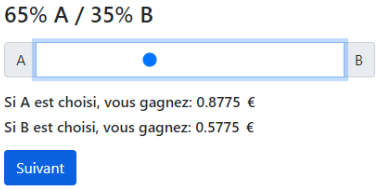
Prediction 1 of 5: What is the probability that you belong to the *mistrusting* vs *trusting* category? Use the slider below.



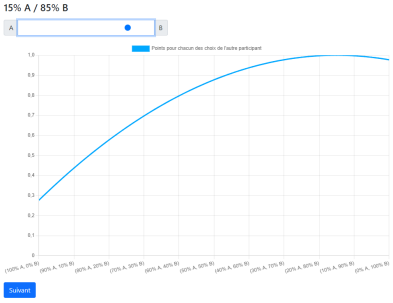
Prediction 2 of 5: What is the probability that a participant classified as *mistrusting* will choose A or B?



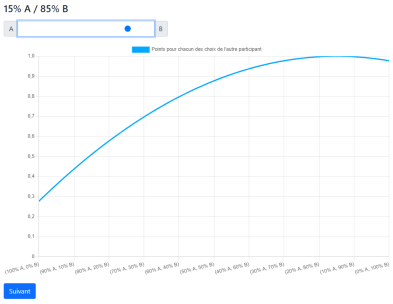
Prediction 3 of 5: What is the probability that a participant classified as *trusting* will choose A or B?



Prediction 4 of 5: What do you believe a mistrusting participant predicts about the probability you will choose A or B?



Prediction 5 of 5: What do you believe a trusting participant predicts about the probability you will choose A or B?



Page 16. Repeated stag-hunt game

You will now begin the main part of the game. You interact in a group of 6 participants over 48 consecutive periods. Pairings within the group are random, but you will encounter an equal number of trusting and mistrusting participants.

In each period, you must:

1. Predict the probability that a participant will choose A or B.
2. Predict the other participant's belief about your decision (i.e., what they think you will choose).

These predictions will allow you to earn payoffs depending on the distance between your prediction and the other participant's actual expectation. Note that for each type of prediction, only one randomly selected period will be paid at the end of the game.

Next, you must choose between A and B in the same game as before.

Your accumulated payoffs across all 48 periods will be added to the payoff from the first game at the beginning of the experiment. At the end of the experiment, you will receive the total payoff from all 49 periods.

Page 17. Instructions adapted to information regimes

Full information
Your group consists of 3 participants categorised as *mistrusting* and 3 participants categorised as *trusting*.
You are in the category *[trusting/mistrusting]*.
In each period, you will be informed of the category of the other participant. Likewise, the other participant will be informed of your category.

Private information
Your group consists of 3 participants categorized as *mistrusting* and 3 participants categorized as *trusting*.
You are in the category *[trusting/mistrusting]*.
Information about the category of other participants will not be revealed to you. The other participants know their own category but have no information about your category or the categories of other participants.

Public information
Your group consists of 3 participants categorized as *mistrusting* and 3 participants categorized as *trusting*.
In each period, you will be informed of the category of the other participant. Likewise, the other participants will be informed of your category.
Information about your own category will not be revealed to you. Likewise, other participants will not have information about their own category.

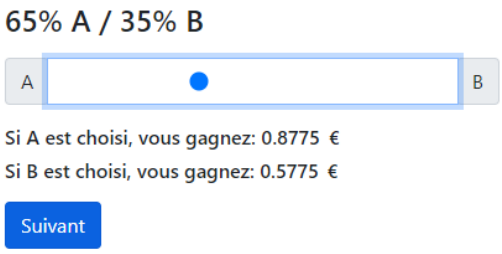
No information
Your group consists of 3 participants categorized as *mistrusting* and 3 participants categorized as *trusting*.

Page 18. FOB (tailored to information regime

You are in the category *[trusting/mistrusting]*.

The other participant in this period is in the category *[trusting/mistrusting]*.

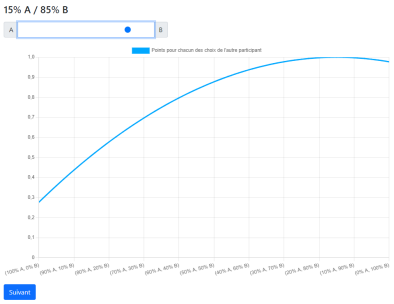
What are the percentage chances that the other participant will choose A and B in this period?



Page 19. SOB (tailored to information regime

The other participant is in the category *[trusting/mistrusting]* and has just estimated the probability percentages you assigned to A and B.

What response do you believe they gave?



Page 20. Stag-hunt decision

You chose *A/B*, the other participant chose *A/B*.

You therefore receive *[gain]* ECUs.

You estimated the probability that the other participant would choose B to be *[FOB]*%. If this period is randomly selected for payment, you may receive €X.

You predicted that the other participant estimated the probability that you would choose B to be *[SOB]*%. The other participant had estimated the probability that you would choose B to be *[SOB]*%. If this period is randomly selected for payment, you may receive €X.

Figure 1.9: Originally registered protocol

How will I know? : The significance of personality awareness for coordination (#131557)

Author(s)
This pre-registration is currently anonymous to enable blind peer-review.
It has 3 authors.

Pre-registered on: 2023/05/09 - 08:09 AM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
We study the impact of information about personality on beliefs and behaviors in a strategic setting (here, a stag-hunt game). The design relies on two personality types, trusting or mistrusting, in repeated interactions where coordination is the most profitable option but is strategically dominated. We analyze these interactions using four environments with varying levels of information given to participants about their own and their counterparts' personalities. Results from previous studies indicate that this type of information leads to the polarization of beliefs and behaviors: when matched with a trusting counterpart, an individual's propensity to coordinate increases; when matched with a mistrusting counterpart, propensity to coordinate decreases. Results also indicate that this pattern is present for individuals with traits from both personalities.

Our results will allow us to test the replicability of the result that knowledge about one's own personality type and the personality type of our counterparts induces a polarization pattern.

The main interest of our study is to identify the four potential micro-mechanisms that may account for the polarization of beliefs and behaviors in this context: (1) perceived similarity (i.e., self-identification with the personality type stereotype), (2) preference-based discrimination, (3) first-order belief-based discrimination (i.e., ostracization of mistrusting personalities), and (4) second-order belief-based discrimination (i.e., anticipation of ostracization of mistrusting personalities by mistrusting personalities). Furthermore, we expand our scope studying whether errors in prediction lead to changes in beliefs and behaviors and if personality-based preferences are altered by repeated interactions.

3) Describe the key dependent variable(s) specifying how they will be measured.
The key dependent variables are beliefs and behaviour in the stag-hunt game:

Choice in the stag-hunt game: this model is a 2x2 simultaneous game with two Nash equilibria in pure strategy with a risky dominant Pareto equilibrium and a non-risky
one. Gains are accumulated.
First-order beliefs: for each stag-hunt game, individuals must predict the choice of their opponent by using a probability slider incentivized through a quadratic scoring rule.
Second-order beliefs: for each stag-hunt game, individuals must predict the first-order beliefs of their opponent by using a probability slider incentivized through a quadratic scoring rule.
Personality types: measuring the score for the Agreeableness trait in a hybrid of two personality tests - HEXACO Personality Inventory and the NEO Personality Inventory (NEO-PI-R) , using three facets from HEXACO (forgiveness, gentleness and flexibility) and a facet from the NEO-PI-R (trust).
Participants respond to a 1-to-5 likert scale, obtain a score and are categorised as trusting if their score is above the median for the session, or mistrusting if their score is below the median for the session.
Five beliefs (beliefs on own personality type, FOB and SOB about a trusting and a mistrusting opponent) are measured twice without incentives, first after the personality questionnaire and then after the repeated stag-hunt game.

4) How many and which conditions will participants be assigned to?
The experimental model consists of four main parts:
A randomised round of a stag-hunt game with first and second-order belief elicitation.
An ex-ante and an ex-post questionnaire to register predictions on personality types and their behaviors.
A personality test to categorise participants as trusting or mistrusting types.
A repeated 48x stag-hunt game with incentivised first and second-order belief elicitation under four treatment conditions.
The four treatment conditions participants are assigned to for part 3 are:
Full Information (FI), meaning information of self and counterpart
Private Information(PrI), meaning information on self but not on counterpart
Public Information (PuI), meaning information on counterpart but not on self
No Information (NI), meaning no information on self nor on counterpart.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Concerning the effect of information about the counterpart's personality on behaviours, the hypotheses will be tested using a logistic regression to predict

probability to cooperate as a function of information in the treatments and beliefs.. Concerning first and second-order beliefs in the stag-hunt game, to study the emergence of mutual beliefs and polarisation, we will run regressions and evaluate the parameters of reinforcement learning models to identify treatments' effects.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We do not plan to remove outliers from the data. We have no exclusion criteria.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Our sample consists of 192 subjects in 8 experimental sessions; each session has 24 participants, 12 for each personality type. The sample size is determined by the necessary amount of registered observations to infer enough significance for the expected probability outcomes of both types of personalities interacting under the four treatment conditions.

Our data set consists then of 9,216 observations in the repeated game, relying on three input variables (first-order beliefs, second-order beliefs and decision) across 48 repeated rounds.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will study how information about personality alters disposition to coordinate through a logistic regression to predict beliefs and behaviour in the first random stag-hunt game. We also include the possibility to further analyse behaviour using the scores in the hybrid test to contrast them against elicited beliefs and the decisions in the repeated game, as well as the impact of information in repeated interactions using the ex antes vs the ex post questionnaires. We will also run the following secondary analyses:
Logistic regression to infer variances in experimental conditions from control to treatment (i.e., No info and PuI to PrI and FI vs PuI to FI).
Logistic regression to predict if awareness of counterpart's personality type affects cooperative decisions: Use FOB, SOB, treatment variables vs counterpart type in each match (i.e., trusting vs mistrusting).
Logistic regression to predict if PuI conditions facilitate the polarisation of mutual beliefs towards cooperation: Use FOB, SOB, self personality type, counterpart personality type and treatment variables (NI vs PuI; PrI vs PuI; FI vs PuI).
Logistic regression to predict if FI and NI conditions facilitate the polarisation of mutual beliefs towards defection: Use FOB, SOB, self personality type, counterpart personality type and treatment variables (NI vs PuI; PrI vs PuI; FI vs PuI).

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4):1115–1153.

Accinelli, E. and Carrera, E. J. (2012). Corruption driven by imitative behavior. *Economics Letters*, 117(1):84–87.

Acedo-Carmona, C. and Gomila, A. (2014). Personal trust increases cooperation beyond general trust. *PLoS ONE*, 9(8):1–10.

Ahloy, J. and Hamman, J. R. (2019). Personality Traits and Endogenous Group Formation. *Source: Revue économique*, 70(6):999–1020.

Akerlof, G. A. and Kranton, R. E. (1997). Social Distance and Social Decisions. Technical Report 5.

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

Akerlof, G. A. and Kranton, R. E. (2002). Identity and Schooling: Some Lessons for the Economics of Education. *Journal of Economic Literature*, 40:1167–1201.

Alfonso-Costillo, A., Brañas-Garza, P., and López-Martín, M. C. (2022). Does the die-under-the-cup device exaggerate cheating? *Economics Letters*, 214.

Amir, A., Kogut, T., and Bereby-Meyer, Y. (2016). Careful cheating: People cheat groups rather than individuals. *Frontiers in Psychology*, 7(371):1–8.

Anvari, F., Wenzel, M., Woodyatt, L., and Haslam, S. A. (2019). The social psychology of whistleblowing: An integrated model. *Organizational Psychology Review*, 9(1):41–67.

Aquino, K. and Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6):1423–1440.

Aron, A., Aron, E., 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):596–612.

Artinger, F., Exadaktylos, F., Koppel, H., and Sääksvuori, L. (2010). Applying Quadratic Scoring Rule transparently in multiple choice settings: A note. *Working Paper*, (January):1–15.

Ashton, M. C. and Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2):150–166.

Ashton, M. C. and Lee, K. (2008). The prediction of Honesty-Humility-related criteria by the HEXACO and Five-Factor Models of personality. *Journal of Research in Personality*, 42(5):1216–1228.

Baader, M., Starmer, C., Tufano, F., and Gächter, S. (2024). Introducing IOS11 as an extended interactive version of the 'Inclusion of Other in the Self' scale to estimate relationship closeness. *Scientific Reports*, 14(1).

Baccini, E. and Hartmann, S. (2022). The Myside Bias in Argument Evaluation: A

Bayesian Model. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022*, pages 1512–1518.

Bäker, A. . and Mechtel, M. (2015). Peer Settings Induce Cheating on Task Performance.

Balliet, D., Wu, J., and De Dreu, C. K. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychological bulletin*, 140(6):1556–1581.

Barranti, M., Carlson, E. N., and Furr, R. M. (2016). Disagreement About Moral Character Is Linked to Interpersonal Costs. *Social Psychological and Personality Science*, 7(8):806–817.

Bartels, D. M. and Burnett, R. C. (2011). A group construal account of drop-in-the-bucket thinking in policy preference and moral judgment. *Journal of Experimental Social Psychology*, 47(1):50–57.

Beer, A. and Watson, D. (2008). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment*, 90(3):250–260.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.

Benistant, J., Galeotti, F., and Villeval, M. C. (2021). The Distinct Impact of Information and Incentives on Cheating The Distinct Impact of Information and Incentives on Cheating *. Technical report.

Benistant, J., Galeotti, F., and Villeval, M. C. (2022). Competition, information, and the erosion of morals. *Journal of Economic Behavior and Organization*, 204:148–163.

Beranek, B. and Castillo, G. (2022). Continuous Inclusion of Other in the Self. Technical report.

Berg, A. (2019). Identity in economics: a review.

Bernard, M., Hett, F., and Mechtel, M. (2016). Social identity and social free-riding. *European Economic Review*, 90:4–17.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, 132:59–72.

Bilancini, E., Boncinelli, L., Capraro, V., Celadin, T., and Di Paolo, R. (2020). "Do the right thing" for whom? An experiment on ingroup favouritism, group assorting and moral suasion. *Judgment and Decision Making*, 15(2):182–192.

Blanken, I., van de Ven, N., and Zeelenberg, M. (2015). A Meta-Analytic Review of Moral Licensing. *Personality and Social Psychology Bulletin*, 41(4):540–558.

Blokland, T. (2012). Blaming neither the undeserving poor nor the revanchist middle classes: A relational approach to marginalization. *Urban Geography*, 33(4):488–507.

Bone, J. E., McAuliffe, K., and Raihani, N. J. (2016). Exploring the motivations for punishment: Framing and country-level effects. *PLoS ONE*, 11(8).

Boone, C., Declerck, C., and Kiyonari, T. (2010). Inducing Cooperative Behavior among Proselfs versus Prosocials: The Moderating Role of Incentives and Trust. *Journal of Conflict Resolution*, 54(5):799–824.

Bose, N. and Sgroi, D. (2022). The role of personality beliefs and "small talk" in strategic behaviour. *PLoS ONE*, 17(9 September).

Brown-Iannuzzi, J. L., Lundberg, K. B., and McKee, S. E. (2021). Economic inequality and socioeconomic ranking inform attitudes toward redistribution. *Journal of Experimental Social Psychology*, 96.

Bussolo, M., Lebrand, M., and Torre, I. (2020). Feeling Poor, Feeling Rich, or Feeling Middle-Class An Empirical Investigation.

Cameron, L., Chaudhuri, A., Erkal, N., and Gangadharan, L. (2009). Propensities to engage in and punish corrupt behavior: Experimental evidence from Australia, India, Indonesia and Singapore. *Journal of Public Economics*, 93(7-8):843–851.

Castillo, G. (2021). Preference reversals with social distances. *Journal of Economic Psychology*, 86.

Castro Santa, J., Exadaktylos, F., and Soto-Faraco, S. (2018). Beliefs about others' intentions determine whether cooperation is the faster choice. *Scientific Reports*, 8(1):1–10.

Chae, J., Kim, K., Kim, Y., Lim, G., Kim, D., and Kim, H. (2022). Ingroup favoritism overrides fairness when resources are limited. *Scientific Reports*, 12(1).

Charness, G. (2000). Self-serving cheap talk: A test of aumann's conjecture. *Games and Economic Behavior*, 33(2):177–194.

Charroin, L., Fortin, B., Villeval, M. C., Boucher, V., Bramoullé, Y., Chen, Y., Cohn, A., Davidson, R., Fluet, C., Marchand, S., and Shearer, B. (2021). Homophily, Peer Effects, and Dishonesty Homophily, Peer Effects, and Dishonesty *. Technical report.

Chierchia, G. and Coricelli, G. (2015). The impact of perceived similarity on tacit coordination: Propensity for matching and aversion to decoupling choices. *Frontiers in Behavioral Neuroscience*, 9(JULY).

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., and Neuberg, S. L. (1997). Reinterpreting the Empathy-Altruism Relationship: When One Into One Equals Oneness. Technical report.

Conrads, J., Irlenbusch, B., Rilke, R. M., and Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, 34:1–7.

Cooper, J. (2019). Cognitive dissonance: Where we've been and where we're going. *International Review of Social Psychology*, 32(1).

Cooper, W. H. and Withey, M. J. (2009). The strong situation hypothesis. *Personality and Social Psychology Review*, 13(1):62–72.

Costa, P. (1992). Neo PI-R professional manual GWAS of Personality View project Lifespan and Intergenerational Effects of Childhood Malnutrition View project. Technical report.

Costa, P. T. and McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment: Volume 2 -*

Personality Measurement and Testing, pages 179–198. SAGE Publications Inc.

Coyne, I. and Bartram, D. (2002). Assessing the Effectiveness of Integrity Tests: A Review. *International Journal of Testing*, 2(1):15–34.

Crede, A.-K. and von Bieberstein, F. (2020). Reputation and lying aversion in the die roll paradigm: Reducing ambiguity fosters honest behavior. *Managerial and Decision Economics*, 41(4).

Crowe, M. L., Lynam, D. R., and Miller, J. D. (2018). Uncovering the structure of agreeableness from self-report measures. *Journal of Personality*, 86(5):771–787.

Currarini, S. and Mengel, F. (2016). Identity, homophily and in-group bias. *European Economic Review*, 90:40–55.

Dalton, R. (2016). Party identification and its implications. *Oxford Research Encyclopedia of Politics*.

de Dreu, C. K. (2010). Social value orientation moderates ingroup love but not outgroup hate in competitive intergroup conflict. *Group Processes & Intergroup Relations*, 13(6):701–713.

De Freitas, J., Thomas, K., DeScioli, P., and Pinker, S. (2019). Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences of the United States of America*, 116(28):13751–13758.

Deutchman, P., Bračič, M., Raihani, N., and McAuliffe, K. (2021). Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evolution and Human Behavior*, 42(1):12–20.

Devetag, G. and Ortmann, A. (2007). When and why? A critical survey on coordination failure in the laboratory. *Experimental Economics*, 10(3):331–344.

Dimant, E. (2019). Contagion of pro- and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73:66–88.

Drouvelis, M. and Georgantzis, N. (2019). Does revealing personality data affect prosocial behaviour? *Journal of Economic Behavior and Organization*, 159:409–420.

Du, H., Chen, A., Chi, P., and King, R. B. (2020). Preprint of "Income Inequality Reduces Civic Honesty".

Dungan, J. A., Young, L., and Waytz, A. (2019). The power of moral concerns in predicting whistleblowing decisions. *Journal of Experimental Social Psychology*, 85.

Easterbrook, M. J., Hadden, I. R., and Nieuwenhuis, M. (2019). Identities in context: How social class shapes inequalities in education. In *The Social Psychology of Inequality*, pages 103–121. Springer International Publishing.

Easterbrook, M. J., Kuppens, T., and Manstead, A. S. (2020). Socioeconomic status and the structure of the self-concept. *British Journal of Social Psychology*, 59(1):66–86.

Elbæk, C. T., Mitkidis, P., Aarøe, L., and Otterbring, T. (2023). Subjective socioeconomic status and income inequality are associated with self-reported morality across 67 countries. *Nature Communications*, 14(1).

Ellemers, N., Pagliaro, S., Barreto, M., and Leach, C. W. (2008). Is It Better to Be Moral

Than Smart? The Effects of Morality and Competence Norms on the Decision to Work at Group Status Improvement. *Journal of Personality and Social Psychology*, 95(6):1397–1410.

Fagbenro, D. A. (2019). Personality Traits and Attitude toward Corruption among Government Workers. *Psychology and Behavioral Science International Journal*, 11(1).

Falk, A. and Zimmermann, F. (2024). Attention and Dread: Experimental Evidence on Preferences for Information. *Management Science*, 70(10):7090–7100.

Fan, C. S., Wei, X., Wu, J., and Zhang, J. (2022). Observability and peer effects: Theory and evidence from a field experiment. *Journal of Economic Behavior and Organization*, 200:847–867.

Fehr, D., Kübler, D., and Danz, D. (2008). Information and Beliefs in a Repeated Normal-form game. *Philosophy of Information*, (3627):551–577.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise-an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Fries, T., Gneezy, U., Kajackaite, A., and Parra, D. (2021). Observability and lying. *Journal of Economic Behavior and Organization*, 189:132–149.

Gächter, S. and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595):496–499.

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).

Galeotti, F., Rilke, R. M., and Verrina, E. (2024). Beliefs and Group Dishonesty: The Role of Strategic Interaction and Responsibility. Technical report.

Gibson, R., Tanner, C., and Wagner Alexander F. (2013). Preferences for Truthfulness: }Heterogeneity Among and Within Individuals. *American Economic Review*, 103(1):532–548.

Gino, F., Ayal, S., and Ariely, D. (2009). Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel. Technical Report 3.

Gino, F., Ayal, S., and Ariely, D. (2012). Self-Serving Altruism? When Unethical Actions That Benefit Others Do Not Trigger Guilt. Technical report.

Giorgetta, C., Grecucci, A., Graffeo, M., Bonini, N., Ferrario, R., and Sanfey, A. G. (2021). Expect the Worst ! Expectations and Social Interactive Decision Making. *Brain Sciences*, 11(572).

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Goldstein, N. J. and Cialdini, R. B. (2007). The spyglass self: A model of vicarious self-perception. *Journal of Personality and Social Psychology*, 92(3):402–417.

Greenberg, S. and Org, C. (2018). Calibration Scoring Rules for Practical Prediction Training. Technical report.

Gries, T., Müller, V., and Jost, J. T. (2022). The Market for Belief Systems: A Formal

Model of Ideological Choice. *Psychological Inquiry*, 33(2):65–83.

Grigoryan, L. (2020). Crossed categorization outside the lab: Findings from a factorial survey experiment. *European Journal of Social Psychology*, 50(5):983–1000.

Grigoryan, L., Seo, S., Simunovic, D., and Hofmann, W. (2023). Helping the ingroup versus harming the outgroup: Evidence from morality-based groups. *Journal of Experimental Social Psychology*, 105.

Gross, J. and De Dreu, C. K. (2021). Rule Following Mitigates Collaborative Cheating and Facilitates the Spreading of Honesty Within Groups. *Personality and Social Psychology Bulletin*, 47(3):395–409.

Gross, J., Leib, M., Offerman, T., and Shalvi, S. (2018). Ethical Free Riding: When Honest People Find Dishonest Partners. *Psychological Science*, 29(12):1956–1968.

Gueguen, N., Jacob, C., and Martin, A. (2009). Mimicry in social interaction: Its effect on human judgment and behavior. *European Journal of Social Sciences*, 8(2):253–259.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M., Lagos, P., Norris, E., Ponarin, and B. Puranen et al. (eds.) (2020). World Values Survey: Round Seven – Country-Pooled Datafile. Technical report, JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Hauge, L. (2007). Identity and Place: A Critical Comparison of Three Identity Theories.

Hauser, O. P., Kraft-Todd, G. T., Rand, D. G., Nowak, M. A., and Norton, M. I. (2021). Invisible inequality leads to punishing the poor and rewarding the rich. *Behavioural Public Policy*, 5(3):333–353.

Hermann, D. and Ostermaier, A. (2018). Be close to me and I will be honest How social distance influences honesty.

Hershcovis, M. S., Neville, L., Reich, T. C., Christie, A. M., Cortina, L. M., and Shan, J. V. (2017). Witnessing wrongdoing: The effects of observer power on incivility intervention in the workplace. *Organizational Behavior and Human Decision Processes*, 142:45–57.

Hewston, M., Rubin, M., and Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, (53):575–604.

Heyman, T., Vankrunkelsven, H., Voorspoels, W., White, A., Storms, G., and Verheyen, S. (2020). When Cheating is an Honest Mistake: A Critical Evaluation of the Matrix Task as a Measure of Dishonesty. *Collabra: Psychology*, 6(1).

Hilbig, B. E., Hessler, C. M., Thielmann, I., Wührl, J., and Zettler, I. (2015). What lies beneath: How the distance between truth and lie drives dishonesty. *Personality and Individual Differences*, 80(2):263–266.

Hilbig, B. E., Zettler, I., Leist, F., and Heydasch, T. (2013). It takes two: Honesty-Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, 54(5):598–603.

Hoffmann, T. (2013). The Effect of Belief Elicitation on Game Play. pages 1–26.

Holt, C. A. and Laury, S. K. (2002). Risk Aversion and Incentive Effects. Technical

Report 5.

Hughes, B. T., Flournoy, J. C., and Srivastava, S. (2020). Is Perceived Similarity More Than Assumed Similarity?: An Interpersonal Path to Seeing Similarity Between Self and Others. *Journal of Personality and Social Psychology*, 121(1):184–200.

Hyndman, K., Terracol, A., and Vaksmann, J. (2013). Beliefs and (In)Stability in Normal-Form Games. (47221).

Inglehart, R. (2000). Culture and Democracy. In *Culture Matters: How Values Shape Human Progress*, pages 80–97. New York: Basic Books.

INSEE Références (2021). La France et ses territoires. Technical report.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2025). The Origins and Consequences of Affective Polarization in the United States. 53:29.

Jansson, F. (2015). What games support the evolution of an ingroup bias? *Journal of Theoretical Biology*, 373:100–110.

Jansson, F. and Eriksson, K. (2015). Cooperation and shared beliefs about trust in the assurance game. *PLoS ONE*, 10(12):1–13.

John, O., Naumann, L., and Soto, C. (2008). *Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues.* Guilford Press, 3rd edition.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90(12).

Jordan, P. J., Troth, A. C., and Yan, H. (2024). Objective and subjective measurement in applied business settings: Improving research in organizations. *Australian Journal of Management*.

Kajonius, P. J. and Dåderman, A. M. (2014). Exploring the relationship between honesty-humility, the big five, and liberal values in Swedish students. *Europe's Journal of Psychology*, 10(1):104–117.

Kaluza, B., Institute, J. S., Kaminka, G., Tambe, M., Kaluža, B., and Kaminka, G. A. (2012). Detection of suspicious behavior from a sparse set of multiagent interactions. Technical report.

Kang, P., Burke, C. J., Tobler, P. N., and Hein, G. (2021). Why we learn less from observing outgroups. *Journal of Neuroscience*, 41(1):144–152.

Kaushik, M., Singh, V., and Chakravarty, S. (2021). Rewards, Detection and Dishonesty: Experimental Evidence from India. *SSRN Electronic Journal*.

Kim, J. E. and Tsvetkova, M. (2021). Cheating in online gaming spreads through observation and victimization. *Network Science*, 9(4):425–442.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms Make Preferences Social. *Journal of the European Economic Association*, 14(3):608–638.

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2019). Intuitive Honesty Versus Dishonesty: Meta-Analytic Evidence. *Perspectives on Psychological Science*, 14(5):778–796.

Kocher, M., Martinsson, P., and Visser, M. (2012). Social background, cooperative behavior, and norm enforcement. *Journal of Economic Behavior and Organization*, 81(2):341–354.

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, 64(9):3995–4008.

Korbel, V. (2016). Do we lie in groups? An experimental evidence. *Applied Economic Letters*, 24(15):1107–1111.

Kraus, M. W., Piff, P. K., and Keltner, D. (2011). Social class as culture: The convergence of resources and rank in the social realm. *Current Directions in Psychological Science*, 20(4):246–250.

Kreps, D. M. (1992). *Game Theory and Economic Modelling*. Oxford.

Kroher, M. and Wolbring, T. (2015). Social control, social learning, and cheating: Evidence from lab and online experiments on dishonesty. *Social Science Research*, 53:311–324.

Ladley, D., Wilkinson, I., and Young, L. (2015). The impact of individual versus group rewards on work group performance and cooperation: A computational social science approach. *Journal of Business Research*, 68(11):2412–2425.

Lane, T. (2023). The strategic use of social identity CeDEx Discussion Paper Series. Technical report.

Larrouy, L. and Lecouteux, G. (2017). Mindreading and endogenous beliefs in games. *Journal of Economic Methodology*, 24(3):318–343.

Le Coq, C., Tremewan, J., and Wagner, A. K. (2015). On the effects of group identity in strategic environments. *European Economic Review*, 76:239–252.

Lee, J. J., Hardin, A. E., Parmar, B., and Gino, F. (2019). The interpersonal costs of dishonesty: How dishonest behavior reduces individuals' ability to read others' emotions. *Journal of Experimental Psychology: General*, 148(9):1557–1574.

Leib, M., Köbis, N., Soraperra, I., Weisel, O., and Shalvi, S. (2021). Collaborative Dishonesty: A Meta-Analytic Review. *Psychological Bulletin*, 147(12):1241–1268.

Leibbrandt, A., López-Pérez, R., and Spiegelman, E. (2023). Reciprocal, but inequality averse as well? Mixed motives for punishment and reward. *Journal of Economic Behavior and Organization*, 210:91–116.

Leidner, B., Castano, E., Zaiser, E., and Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin*, 36(8):1115–1129.

Lönnqvist, J. E., Ilmarinen, V. J., and Verkasalo, M. (2021). Who likes whom? The interaction between perceiver personality and target look. *Journal of Research in Personality*, 90.

Loustau, T., Glassman, J., Martin, J. W., Young, L., and McAuliffe, K. (2024). The impact of group membership on punishment versus partner rejection. *Scientific Reports*, 14(1).

Lutz, G. and Lauener, L. (2020). Measuring party affiliation. Technical report, Lausanne: Swiss Centre of Expertise in the Social Sciences (FORS)., Lausanne.

Macků, K., Caha, J., Pászto, V., and Tuček, P. (2020). Subjective or objective? How objective measures relate to subjective life satisfaction in Europe. *ISPRS International Journal of Geo-Information*, 9(5).

Magni, G. (2021). Economic inequality, immigrants and selective solidarity: From perceived lack of opportunity to in-group favoritism.

Mann, H., Garcia-Rada, X., Houser, D., and Ariely, D. (2014). Everybody else is doing it: Exploring social transmission of lying behavior. *PLoS ONE*, 9(10).

Manstead, A. S. (2018). The psychology of social class: How socioeconomic status impacts thought, feelings, and behaviour. *British Journal of Social Psychology*, 57(2):267–291.

Martin, R. A. (2015). *Perceived and Actual Similarity as Predictors of Self-Disclosure and Perceived Understanding at Zero Acquaintance*. PhD thesis.

Martinangeli, A. F. and Martinsson, P. (2020). We, the rich: Inequality, identity and cooperation. *Journal of Economic Behavior and Organization*, 178:249–266.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people. *Journal of Marketing Research*, 45(6):633–644.

McFerran, B., Aquino, K., and Duffy, M. (2010). How Personality and Moral Identity Relate to Individuals' Ethical Ideology. *Business Ethics Quarterly*, 20(1):35–56.

Mendoza, S. A., Lane, S. P., and Amodio, D. M. (2014). For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game. *Social Psychological and Personality Science*, 5(6):662–670.

Meyners, J., Barrot, C., Becker, J. U., and Goldenberg, J. (2017). The role of mere closeness: How Geographic proximity affects social influence. *Journal of Marketing*, 81(5):49–66.

Michaeli, M. (2020). Grouping, in-group bias and the cost of cheating. *Games and Economic Behavior*, 121:90–107.

Molho, C., De Petrillo, F., Garfield, Z. H., and Slewe, S. (2024). Cross-societal variation in norm enforcement systems.

Moss, R. H., Kelly, B., Bird, P. K., and Pickett, K. E. (2023). Examining individual social status using the MacArthur Scale of Subjective Social Status: Findings from the Born in Bradford study. *SSM - Population Health*, 23.

OECD (2024). OECD Survey on Drivers of Trust in Public Institutions – 2024 Results: Building Trust in a Complex Policy
}Environment. Technical report, OECD Publishing, Paris.

Offerman, T., Sonnemans, J., van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies*, 76(4):1461–1489.

Owuamalam, C. K., Rubin, M., Spears, R., and Weerabangsa, M. M. a. (2017). Why Do

People from Low-Status Groups Support Class Systems that Disadvantage Them? A Test of Two Mainstream Explanations in Malaysia and Australia. *Journal of Social Issues*, 73(1):80–98.

Panagopoulos, C., Leighley, J. E., and Hamel, B. T. (2017). Are Voters Mobilized by a 'Friend-and-Neighbor'on the Ballot? Evidence from a Field Experiment. *Political Behavior*, 39(4):865–882.

Pansini, R., Campennì, M., and Shi, L. (2018). Asymmetric use of punishment in socioeconomic segregated societies leads to an unequal distribution of wealth. Technical report.

Proto, E., Rustichini, A., Deyoung, C., Friebel, G., Grimalda, G., Isoni, A., Loomes, G., Manzini, P., Mariotti, M., Miller, J., Oswald, A., and Stewart, N. (2014). Cooperation and Personality. Technical report.

Pulfrey, C., Durussel, K., and Butera, F. (2018). The good cheat: Benevolence and the justification of collective cheating. *Journal of Educational Psychology*, 110(6):764–784.

Rantakari, H. (2023). How to reward honesty? *Journal of Economic Behavior and Organization*, 207:129–145.

Régner, I. and Monteil, J.-M. (2007). Low-and high-socioeconomic status students preference for ingroup comparisons and their underpinning ability expectations. *Revue Internationale de Psychologie Sociale*, 20(1):87–104.

Renger, D., Lohmann, J. F., Renger, S., and Martiny, S. E. (2024). Socioeconomic status and self-regard income predicts self-respect over time. *Social Psychology*, 55(1):12–24.

Rijnks, R. H. and Strijker, D. (2013). Spatial effects on the image and identity of a rural area. *Journal of Environmental Psychology*, 36:103–111.

Robalo, P., Schram, A., and Sonnemans, J. (2017). Other-regarding preferences, in-group bias and political participation: An experiment. *Journal of Economic Psychology*, 62:130–154.

Rothstein, B. (2011). Anti-corruption: The indirect 'big bang' approach. *Review of International Political Economy*, 18(2):228–250.

Rothstein, B. and Eek, D. (2009). Political Corruption and Social Trust. *Rationality and Society*, 21(1):81–112.

Rubin, M., Badea, C., and Jetten, J. (2014). Low status groups show in-group favoritism to compensate for their low status and compete for higher status. *Group Processes & Intergroup Relations*, 17(5):563–576.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under "almost common knowledge". *American Economic Review*, 79(3):385–391.

Rubinstein, A. and Salant, Y. (2016). Isn't everyone like me?": On the presence of self-similarity in strategic interactions. *Judgment and Decision Making*, 11(2):168–173.

Ruch, W., Bruntsch, R., and Wagner, L. (2017). The role of character traits in economic

games. *Personality and Individual Differences*, 108:186–190.

Rullo, M., Monaco, S., Giannini, F., Livi, S., and Presaghi, F. (2019). In the name of truth: People's reactions to ingroup and outgroup members who self-disclose a severe error. *Social Science Journal*, 56(3):421–424.

Rullo, M., Presaghi, F., Baldner, C., Livi, S., and Butera, F. (2024). Omertà in intragroup cheating: The role of ingroup identity in dishonesty and whistleblowing. *Group Processes and Intergroup Relations*, 27(1):41–61.

Rustichini, A. (2009). Neuroeconomics: what have we found, and what should we search for.

Rustichini, A., DeYoung, C. G., Anderson, J. E., and Burks, S. V. (2016). Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation. *Journal of Behavioral and Experimental Economics*, 64:122–137.

Ruzzier, C. A. and Woo, M. D. (2023). Discrimination with inaccurate beliefs and confirmation bias. *Journal of Economic Behavior and Organization*, 210:379–390.

Ryvkin, D., Serra, D., and Tremewan, J. (2017). I paid a bribe: An experiment on information sharing and extortionary corruption. *European Economic Review*, 94:1–22.

Schiller, B., Baumgartner, T., and Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior*, 35(3):169–175.

Schram, A., Zheng, J. D., and Zhuravleva, T. (2022). Corruption: A cross-country comparison of contagion and conformism. *Journal of Economic Behavior and Organization*, 193:497–518.

Sgroi, D., Yeo, J., and Zhuo, S. (2021). Ingroup Bias with Multiple Identities: The Case of Religion and Attitudes Towards Government Size. Technical report.

Shalvi, S., Dana, J., Handgraaf, M. J., and De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190.

Siniver, E., Tobol, Y., and Yaniv, G. (2022). Collective Punishment and Cheating in the Die-Under-the-Cup Task. *Experimental Psychology*, 69(1):40–45.

Skyrms, B. (2003). *The Stag Hunt and the Evolution of Social Structure*. Number 1. Cambridge University Press, Cambridge.

Sosa, M. and Maoret, M. (2023). Close to Me: The Impact of the Interplay of Physical and Social Proximity on Dyadic Collaboration Effectiveness. Technical report.

Stahl, D. and Huyck, J. V. (2002). Learning conditional behavior in similar stag hunt games. (January).

Steinel, W., Valtcheva, K., Gross, J., Celse, J., Max, S., and Shalvi, S. (2022). (Dis)honesty in the face of uncertain gains or losses. *Journal of Economic Psychology*, 90.

Tajfel, H. and Turner, J. C. (1979). An integrative theory of intergroup conflict. In *The social psychology of intergroup relations*, pages 33–37. Brooks/Cole, Monterey, CA.

Thielmann, I., Akrami, N., Babarović, T., Belloch, A., Bergh, R., Chirumbolo, A., Čolović, P., de Vries, R. E., Dostál, D., Egorova, M., Gnisci, A., Heydasch, T., Hilbig, B. E., Hsu, K. Y., Izdebski, P., Leone, L., Marcus, B., Međedović, J., Nagy, J., Parshikova, O., Perugini, M., Petrović, B., Romero, E., Sergi, I., Shin, K. H., Smederevac, S., Šverko, I., Szarota, P., Szirmák, Z., Tatar, A., Wakabayashi, A., Wasti, S. A., Záškodná, T., Zettler, I., Ashton, M. C., and Lee, K. (2020). The HEXACO–100 Across 16 Languages: A Large-Scale Test of Measurement Invariance. *Journal of Personality Assessment*, 102(5):714–726.

Thielmann, I., Hilbig, B. E., Klein, S. A., Seidl, A., and Heck, D. W. (2024). Cheating to benefit others? On the relation between Honesty-Humility and prosocial lies. *Journal of Personality*, 92(3):870–882.

Thomas, G. O., Poortinga, W., and Sautkina, E. (2016). The Welsh Single-Use Carrier Bag Charge and behavioural spillover. *Journal of Environmental Psychology*, 47(2880):126–135.

Thomas, K. A., DeScioli, P., Haque, O. S., and Pinker, S. (2014). The Psychology of Coordination and Common Knowledge. *Journal of Personality and Social Psychology*, 107(4):657–676.

Tobol, Y., Siniver, E., and Yaniv, G. (2020). Do tightwads cheat more? Evidence from three field experiments. *Journal of Economic Behavior and Organization*, 180:148–158.

Tsvetkova, M. and Macy, M. W. (2015). The social contagion of antisocial behavior. *Sociological Science*, 2:36–49.

Van Assche, J., Politi, E., Van Dessel, P., and Phalet, K. (2020). To punish or to assist? Divergent reactions to ingroup and outgroup members disobeying social distancing. *British Journal of Social Psychology*, 59(3):594–606.

van de Ven, J. and Villeval, M. C. (2015). Dishonesty under scrutiny. *Journal of the Economic Science Association*, 1(1):86–99.

Van De Walle, S. (2008). *Perceptions of corruption as distrust? Cause and effect in attitudes toward government*. Number June.

Van Huyck, J., Viriyavipart, A., and Brown, A. L. (2018). When less information is good enough: experiments with global stag hunt games. *Experimental Economics*, 21(3):527–548.

Van Huyck, J. B., Battalio, R. C., and Beil, R. O. (1990). Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure. *American Economic Review*, 80(1):234–248.

van Oosten, S. (2025). The Importance of In-group Favoritism in Explaining Voting for PRRPs: A Study of Minority and Majority Groups in France, Germany and the Netherlands. Technical report, European Center for Populism Studies, Brussels.

Volk, S., Thöni, C., and Ruigrok, W. (2011). Personality, personal values and cooperation preferences in public goods games: A longitudinal study. *Personality and Individual Differences*, 50(6):810–815.

Waytz, A., Dungan, J., and Young, L. (2013). The whistleblower's dilemma and the fairness-loyalty tradeoff. *Journal of Experimental Social Psychology*, 49(6):1027–1033.

Weiner, D. S. and Laurent, S. M. (2021). The (Income-Adjusted) Price of Good Behavior: Documenting the Counter-Intuitive, Wealth-Based Moral Judgment Gap. *Journal of Experimental Psychology: General*, 150(3):484–506.

Weisel, O. and Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences of the United States of America*, 112(34):10651–10656.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, 115(2):157–168.

Windrich, I., Kierspel, S., Neumann, T., Berger, R., and Vogt, B. (2024). Enforcement of Fairness Norms by Punishment: A Comparison of Gains and Losses. *Behavioral Sciences*, 14(1).

Winter, F. and Zhang, N. (2018). Social norm enforcement in ethnically diverse communities. 115(11):2722–2727.

Wu, J., Balliet, D., and Van Lange, P. A. (2016). Reputation, Gossip, and Human Cooperation.

Zhao, K. and Smillie, L. D. (2015). The Role of Interpersonal Traits in Social Decision Making: Exploring Sources of Behavioral Heterogeneity in Economic Games. *Personality and Social Psychology Review*, 19(3):277–302.

Zhou, L., Su, C., Sun, X., Zhao, X., and Choo, K. K. R. (2018). Stag hunt and trust emergence in social networks. *Future Generation Computer Systems*, 88:168–172.

Chapter 2. How close is close enough? When social closeness backfires on honesty

Irving Argaez Corona[a], Béatrice Boulu-Reshef[b], Jean-Christophe Vergnaud[a,c]

[a]Centre d'Économie de la Sorbonne (CES), Université Paris 1 Panthéon-Sorbonne, France
[b]CY Cergy Paris Université (THEMA), France
[c]Centre National de la Recherche Scientifique (CNRS), France

Abstract

The relationship between dishonesty and social closeness has garnered increasing attention from scholars. While the literature has long evidenced that social closeness increases cooperation, recent work suggests it may also enable cheating behaviour through in-group justification. We study this relationship in an online Die-under-the-cup task (DUTC), asking whether misreporting outcomes increases when participants are paired with socially close rather than socially distant counterparts. We recruited 288 participants and implemented two treatments that made social closeness salient along socioeconomic status (T1) and political alignment (T2). We modelled closeness objectively (living in localities with comparable socioeconomic levels and administered by the same political party), as well as subjectively (self-reported personal income and political preferences matching locality averages). Across pooled and treatment-specific analyses, we find little evidence that social closeness systematically increases misreporting in the DUTC, as differences in reported payoffs are small and sensitive to specification. While objective distance shows weak and non-robust associations with behaviour, subjective measures of closeness are consistently non-significant. Furthermore, we also examine whether being observed by a socially close counterpart amplifies misreports and do not detect a reliable effect, aside from isolated, non-generalisable patterns. Our results suggest that any relationship between social closeness and cheating behaviour in the DUTC is limited and context-dependent. Our findings underscore the importance of multi-method measurement when evaluating how social closeness relates to strategic decision-making.

This work has been funded by a French government subsidy managed by the *Agence Nationale de la Recherche* under the framework of the *Investissements d'avenir* programme, reference **ANR-17-EURE-001**. The paper received approval from the Institutional Review Board of Paris School of Economics (decision 2024-05). The pre-registered research protocol can be found in this link for peer-review purposes.

Résumé

La relation entre malhonnêteté et proximité sociale suscite un intérêt croissant parmi les chercheurs. Si la littérature montre depuis longtemps que la proximité sociale favorise la coopération, des travaux récents suggèrent qu'elle peut également faciliter des comportements non éthiques via des formes de justification intra-groupe. Nous étudions cette relation à l'aide d'une tâche virtuelle de « Die-under-the-cup, DUTC », en nous demandant si la triche augmente lorsque les participants sont appariés à des homologues socialement proches plutôt que socialement éloignés. Nous avons recruté 288 participants et mis en place deux traitements rendant saillante la proximité sociale selon le statut socio-économique (T1) et l'alignement politique (T2). Nous avons modélisé la proximité de manière objective (résider dans des localités de niveau socio-économique comparable et administrées par le même parti politique), ainsi que de manière subjective (revenu personnel autodéclaré et préférences politiques en adéquation avec les moyennes de la localité). Dans l'ensemble des analyses, qu'elles soient regroupées ou spécifiques à chaque traitement, nous trouvons peu d'éléments indiquant que la proximité sociale augmente systématiquement la tricherie : les différences de gains déclarés sont faibles et sensibles aux spécifications retenues. Alors que la distance objective présente des associations faibles et peu robustes avec le comportement, les mesures subjectives de proximité ne sont jamais significatives. Nous examinons également si le fait d'être observé par un pair socialement proche amplifie la tricherie et ne détectons pas d'effet robuste, hormis quelques schémas isolés et non généralisables. Nos résultats suggèrent que la relation entre proximité sociale et comportement malhonnête dans la DUTC est limitée et dépend fortement du contexte. Ils soulignent l'importance de recourir à des mesures multiméthodes pour évaluer la manière dont la proximité sociale se rattache à la prise de décision stratégique.

2.1 Introduction

Unethical behaviour has long been a subject of academic interest within the fields of behavioural sciences, economics and social psychology. A growing body of research suggests that these behaviours are not merely individual acts driven by personal gain, but a social phenomenon enabled by group dynamics and social contexts (Bäker and Mechtel, 2015; Goldstein and Cialdini, 2007; Hermann and Ostermaier, 2018; Rullo et al., 2024; **?**). However, the specific mechanisms that enhance unethical behaviour in close social contexts remain underexplored. This paper seeks to narrow this gap with a two-fold purpose: to examine whether the propensity to cheat is associated with social closeness and whether observability by a socially close counterpart facilitates unethical behaviour.

The literature on dishonesty consistently demonstrates that individuals act unethically to the extent that they can justify this behaviour to themselves and to others (Fischbacher and Föllmi-Heusi, 2013; Gächter and Schulz, 2016; Mazar et al., 2008). Lacking the mechanisms to justify this behaviour can deter individuals from dishonesty, as they constantly strive to maintain a positive image within their social circles (Amir et al., 2016; Barranti et al., 2016). Evidence also suggests that portraying oneself as a moral person reduces dishonesty when one is uncertain about the corruptibility of a counterpart and when the costs of being dishonest are elevated (Mendoza et al., 2014; Ryvkin et al., 2017; Wu et al., 2016). A meta-analysis on preferences for truth-telling led by Abeler et al. (2019) concluded that aversion to lying increases when both, reputational costs and social norms, are at stake.

In recent years the research has ventured into explaining some of the mechanisms that link social closeness to unethical behaviour. These include social cues such as in-group bias, peer observability or collaborative dishonesty (Abeler et al., 2019; Fan et al., 2022; Fries et al., 2021; Grigoryan, 2020; Mann et al., 2014). Some key findings point to lack of accountability (Bäker and Mechtel, 2015; Fan et al., 2022), in-group benefits (Pulfrey et al., 2018; Rullo et al., 2024) and social closeness to rule violators as strong enablers of unethical behaviour (Gino et al., 2012; Gross et al., 2018) .

In this paper, we revisit these questions in an online Die-under-the-cup (DUTC) task that allows participants to cheat in their payoffs. Our design makes social closeness salient along two dimensions: a socioeconomic treatment (Treatment 1) and a political treatment (Treatment 2). Social closeness is operationalised with an objective measure: whether participants live in localities with comparable income levels (T1) or elected deputies belonging to the same political party (T2); as well as a subjective measure: whether participants self-reported income corresponds to the average income level in their department (T1), or whether their self-reported political party preferences correspond to the elected party in their department of residence (T2).

Our model uses reported payoffs across 24 rounds of the DUTC as a standard behavioural proxy for cheating. We analyse the data in three complementary ways: first, we run simple mean comparisons between close and distant pairings to gauge differences. Second, we estimate OLS models with robust standard errors to test effects and swap in alternative measures of distance (binary and three-level; for politics also a continuous specification) to test robustness. Third, we compute within-participant changes in average payoffs across social observation to isolate shifts associated with moving from close to distant observation for the same person.

Across these analyses, results indicate no consistent evidence that social closeness meaningfully alters cheating. Any directional patterns are modest and do not persist across treatments, pairing types, or alternative ways of defining closeness. Socioeconomic and political settings yield similar, inconclusive patterns; objective and subjective measures do not show systematic differences; and within-participant comparisons do not reveal stable shifts when individuals move from close-to-distant observation. Likewise, being observed by a socially close counterpart does not reliably amplify cheating. Furthermore, our results suggest that any relationship between social closeness and misreporting in this setting is limited and context-dependent, underscoring the importance of careful measurement and design when evaluating how social ties shape dishonest decision-making.

The paper proceeds as follows: Section 2.2 elaborates on the existing literature. In section 2.3 we present the experimental design. Section 2.4 reports the descriptive statistics, while section 2.5 reports the results. Section 2.6 concludes.

2.2 Theoretical background and related literature

2.2.1 The socialisation and justification of cheating

Cheating involves both cognitive work and norm violation. Individuals weigh gains against the mental effort of self-justification and the reputational costs of appearing dishonest (Abeler et al., 2019; Barranti et al., 2016; Cameron et al., 2009; Fischbacher and Föllmi-Heusi, 2013; Gächter and Schulz, 2016; Shalvi et al., 2011). To avoid these costs, people often construct self-serving rationales, facilitating the diffusion of dishonest behaviour within networks (Accinelli and Carrera, 2012; Hilbig et al., 2013; Mazar et al., 2008; Schram et al., 2022; Weisel and Shalvi, 2015).

The literature highlights certain circumstances that facilitate cheating: ignoring the identity of those affected by one's dishonesty (Bénabou and Tirole, 2016; Gneezy, 2005; Köbis et al., 2019); targeting groups rather than identifiable individuals (Amir et al., 2016; Bartels and Burnett, 2011); or being overconfident about not getting caught (Gibson et al., 2013; Mendoza et al., 2014). Similarly, the literature underpins that certain socialisation mechanisms legitimise cheating, including collaborative cheating and in-group favouritism (Gino et al., 2009; Köbis et al., 2019; Lee et al., 2019; Thielmann et al., 2024; Weisel and Shalvi, 2015). Evidence shows that cheating increases when the benefits are shared and when there are permissive attitudes toward the cheating of in-group members (Pulfrey et al., 2018; Wiltermuth, 2011).

Work on social identity also documents robust in-group preference Balliet et al. (2014); de Dreu (2010); Hewston et al. (2002); Le Coq et al. (2015); Michaeli (2020), however, the root causes of this type of behaviour and how relatedness comes to matter in pure, low-stake environments remain underexplored. In this paper, we address this gap by isolating social closeness as a treatment variable, using exogenous information on socio-economic status and political alignment to vary objective and subjective similarity between participants and passive observers in a Die-under-the-cup task where misreporting affects only personal gains.

2.2.2 Social closeness and observability in experimental paradigms

This paper sits at the intersection of two streams of literature: work using the DUTC game to infer dishonest behaviour and work on how social closeness and identity shape dishonesty. While both fields have been explored independently, research integrating these perspectives remains limited.

Previous work shows that feedback and strategic interaction are central in shaping cheating in the Die-under-the-cup paradigm (DUTC). Kroher and Wolbring (2015) compare lab and online implementations with varying feedback, finding that paired participants gradually coordinate on higher reports despite initially lower cheating relative to controls. Siniver et al. (2022) study collective punishment and observe the opposite of canonical deterrence: cheating increases when punishment is introduced, as participants accept penalties if group benefits are at stake. In a 24-round design, Benistant et al. (2021) show that cheating rises under tournament incentives (versus piece-rate), when opportunities to cheat are present (versus absent), and under continuous feedback (versus delayed ex post), underscoring how incentives and information dynamics amplify misreporting.

Closer to our question, Hermann and Ostermaier (2018) tie DUTC payoffs to a subsequent Dictator allocation and manipulate social distance via the recipient (experimenter versus fellow students), finding higher cheating under greater distance. Beyond the DUTC task, identity-based proximity shapes both wrongdoing and its policing. Rullo et al. (2024) show that shared nationality increases cheating and reduces whistleblowing in an investment task, with mediation by cheating; Bicchieri et al. (2022) find that peer observability reduces norm compliance and that social proximity strengthens this effect in a lab donation game using minimal groups. Overall, these results indicate that social ties can legitimise or shield norm violations and that social distance modulates third-party responses.

Recent work also nuances the role of observability and responsibility. van de Ven and Villeval (2015) show that putting deceptive messages under the scrutiny of a third party who observes payoffs and can even denounce lies has, at best, a limited impact on dishonesty. In group-dishonesty settings, Galeotti et al. (2024) find that beliefs about partners' dishonesty and the strategic structure of payoffs are key determinants of lying, whereas sharing responsibility only modestly affects behaviour for a subset of participants. These findings suggest that how individuals anticipate others' behaviour and how payoffs are tied across group members plays a larger role in shaping dishonesty than scrutiny or shared responsibility alone.

In this sense, our paper takes a step back to ask whether, in environments with independent payoffs, no cheating signals bor formal sanctions, behaviour changes depending on closeness to the Observer. We isolate social closeness from strategic interdependence by studying a repeated DUTC in which Die-rollers' reports determine only their own earnings and observers cannot affect payoffs. Within this environment, we implement real-world measures of social proximity, combining department-level socioeconomic tiers and political alignment with subjective self-placement, to vary both objective and perceived similarity between Die-rollers and Observers. This design allows us to test whether socioeconomic status and political alignment are reliable mediators of closeness in explaining cheating when one is watched by a socially close versus distant counterpart. In doing so, we move beyond minimal-group manipulations and settings

with strong payoff ties, and provide evidence on whether social closeness alone is sufficient to distort behaviour.

2.2.3 Contributions and Hypotheses

Our paper advances the literature with three contributions:

(1) A multi-scale operationalisation of social closeness. We build parallel objective and subjective measures of closeness along two real-world dimensions, socioeconomic status (T1) and political alignment (T2). Objective closeness is based on department-level information; subjective closeness maps individuals' self-reported income and party preferences onto the dominant income tier and elected party in their department. For each dimension, we first define a binary indicator of closeness (1 if both live in departments from same income tier or elected a deputy from the same party; 0 otherwise). We then refine this into an ordered distance measure, distinguishing short distance (e.g., adjacent SES tiers: low/medium, medium/high; or adjacent political positions: left/centre-right, centre-right/far-right) from long distances (e.g., low/high SES; left/far-right). This multi-scale approach allows us to test whether the magnitude of distance moderates behaviour and to compare effects across alternative operalisation of social closeness.

(2) A two-block, within-participant observation design. Each participant is observed in two blocks that differ only in the counterpart's social closeness: one socially close and one socially distant observer. Because the observer's presence does not alter payoffs or impose penalties/rewards, any differences in reports across blocks can be attributed to observability under social closeness rather than strategic enforcement or self-interested sanctioning.

(3) Socioeconomic and political selectivity in pairings. We then test whether the same behavioural proxy responds differently across socioeconomic vs. political contexts and whether ordered distance (short vs. long) moderates any association. This design speaks directly to whether selectivity in suspicions of cheating relate to real-world social boundaries.

Against this background and guided by the existing literature, we test the following hypotheses:

Hypothesis 1: Social closeness is associated with higher propensity to cheat in outcome reports, whether modelled objectively or subjectively.

Hypothesis 2: Cheating in outcome reports is higher when participants are observed by socially close counterparts.

2.3 Experimental design

Our experimental setting consisted of three tasks: (1) a socioeconomic survey to register participants' socioeconomic, political and demographic data, a (2) a Die-under-the-cup (DUTC) task to be played for 24 rounds in rotating pairs, and a (3) the Inclusion of the Other in the Self (IOS) scale to measure perceptions of closeness among participants. The experiment had two

treatment conditions: socioeconomic status (T1) and political alignment (T2). In the following sections we provide a detailed overview of the experimental design and its tasks.

Task 1. Socioeconomic Questionnaire

Participants completed a socioeconomic questionnaire that collected demographic data (age, gender, marital status, scholarity and education) and socioeconomic data - monthly income, partner's income (if applicable), weekly working hours, accommodation status (e.g., homeowner, tenant, free lodging), current credit situation, and social security benefits received (if any). Given that a significant portion of participants were students, we adapted the income question to include: *If you are part of your family's tax household, what is the approximate net monthly household income?*.

The survey included a question on political alignment: *On a scale from 0 to 10, where 0 represents "Strongly disagree" and 10 represents "Strongly agree", what is your level of agreement with the policies of these political parties?* followed by three sliders to rate their level of agreement with the three major political forces in France, as per the results from the 2024 legislative elections (i.e., the left-wing coalition *Nouveau Front Populaire* (NFP), the centre-right presidential alliance *Ensemble pour la République* (EPR), and the far-right party and its allies *Rassemblement National* (RN)).

Task 2. Online Die-under-the-cup task (DUTC)

Participants complete 24 consecutive rounds of the DUTC in fixed, randomly assigned roles:

- **Die-roller (P1):** privately rolls a die, sees the true outcome and registers a value for payment.
- **Observer (P2):** sees the true die outcome and P1's registered value, then registers their own payoff.

Payoffs equal the registered die value in euros, except for a roll of 6 which yields 0€. With a fair die each outcome occurs with probability 1/6, so under truthful reporting the expected mean is

$$\frac{1+2+3+4+5+0}{6} \;=\; 2.5.$$

One round among the 24 is randomly selected for payment at the end.

Each round follows this sequence:

1. **P1 private roll.** A virtual die appears on P1's screen. When P1 clicks *Roll*, the outcome is generated locally using the browser's cryptographic randomness, and the true die face $k \in \{1, \ldots, 6\}$ is shown to P1.
2. **P1 registers a value.** P1 enters the value to be used for payment.
3. **P2 observes and reports.** P2 sees the true die outcome and P1's report for that round, and registers their own payoff.

4. **Data handling.** After each round, the true die value is not retained in the research dataset. The server logs store only session metadata (session ID, round number, treatment, player role) and the values registered for payment. True die roll outcomes are not accessible to experimenters *ex post*, ensuring that cheating can only be inferred statistically from suspicious report distributions.

Importantly, P1 never observes P2's reported payoff, nor receives any feedback about P2's behaviour over the 24 rounds. Both participants' payoffs depend solely on their own reports and are independent of each other. P2 thus provides *passive observability* of P1's report without creating channels for strategic coordination or feedback-based social influence.

Our design also manipulates social closeness in observation: each participant completes two 12-round blocks that differ only in the social relation between P1 and P2, with one block paired with a socially close observer and the other with a socially distant observer (the order of blocks is randomised across participants).

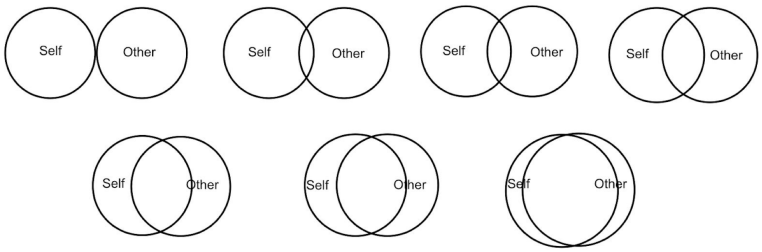
First introduced in Fischbacher and Föllmi-Heusi (2013) and later adapted online by Kocher et al. (2018), the DUTC has been widely used for detecting cheating in controlled environments. This setting provides a suitable framework for our study, as it offers anonymity and incentivises dishonest reporting through monetary rewards based on self-reported outcomes (Benistant et al., 2022; Siniver et al., 2022; Steinel et al., 2022). The DUTC has proven a useful experimental setting to test different mechanisms related to cheating, such as corrupt collaboration (Weisel and Shalvi, 2015), collective punishment (Siniver et al., 2022), reputational costs (Crede and von Bieberstein, 2020) and favourable self-concept maintenance (Gino et al., 2012).

Task 3. Inclusion of the Other in the Self (IOS) Scale

As a robustness check on the social closeness mechanisms, we measured participants' perceived closeness using the Inclusion of IOS Scale (see Aron et al. (1992)). The scale consists of seven pairs of circles, one labelled "self" and one labelled "other," varying in overlap from no overlap (no closeness) to full overlap (strong closeness), see Figure 2.10. At the end of the experiment (*ex-post* the 24-round DUTC), participants registered two things: (a) their feelings of closeness to counterparts from their own department `ios_same` and (b) their feelings of closeness to counterparts from different departments `ios_different`. We adapted the original pictorial representation into French, replacing "self" with "you," following Baader et al. (2024).

The IOS scale has been widely validated as a reliable measure of intimacy and emotional closeness (Beranek and Castillo, 2022; Gächter et al., 2015). Studies also demonstrate its stability over time (Baader et al., 2024; Cialdini et al., 1997) and its sensitivity to social distance in shaping personal preferences (Castillo, 2021).

Figure 2.10: IOS Scale



Treatment conditions

At the recruitment stage, the only information available about participants was their postal code, which we used to identify their department of residence. Based on department-level socioeconomic data and the political affiliation of the deputy elected in their department, we invited participants to secure a balanced between-group variation. We implemented two treatment conditions.

In Treatment 1 (socioeconomic status), participants were classified as residents of *low*, *medium* or *high*-income departments. Each T1 session was built around two income tiers at a time (e.g. Low and Medium; Medium and High). When 48 participants had logged in, the session consisted of 24 Die-rollers (P1) and 24 Observers (P2), with half of each role drawn from each of the two income tiers. Within these sessions, pairs were coded as socially close when both participants resided in departments from the same tier, and as socially distant when they belonged to different tiers.

In Treatment 2 (political alignment), participants were classified according to the party affiliation of the elected deputy in their department. We based this decision on the 2024 French legislative elections, where three major political forces contended: *Nouveau Front Populaire* (NFP), *Ensemble pour la République* (EPR), and *Rassemblement National* (RN). Each T2 session was similarly built around two party blocs at a time (e.g. NFP and EPR; EPR and RN), again with 48 participants (24 P1 and 24 P2) and half of each role drawn from each bloc. Within these sessions, pairs were coded as socially close when participants lived in departments represented by deputies from the same political bloc, otherwise as socially distant .

In both treatments, the pairing algorithm ensured that each participant faced an equal number of socially close and socially distant counterparts across the 24 rounds (see Table 2.13). Once a treatment session for a given pair of SES tiers or political categories (e.g. Low–Medium or NFP–EPR) was filled with 48 participants, further volunteers were allocated to an independent play group, along with other participants whose department did not meet the specific criteria for either treatment. This independent group served to secure an even distribution of experimental sessions in the treatments, however, its observations were not used for analysis. Appendix Table 2.20 provides a detailed breakdown of the departments used in the sample.

Table 2.13: Session composition by treatment and social closeness

T1 Socioeconomic treatment			T2 Political treatment		
Session	**P1 type**	**P2 types in session**	**Session**	**P1 type**	**P2 types in session**
Low–Medium	Low income	Low (SC), Medium (SD)	NFP–EPR	NFP	NFP (SC), EPR (SD)
	Medium income	Medium (SC), Low (SD)		EPR	EPR (SC), NFP (SD)
Low–High	Low income	Low (SC), High (SD)	NFP–RN	NFP	NFP (SC), RN (SD)
	High income	High (SC), Low (SD)		RN	RN (SC), NFP (SD)
Medium–High	Medium income	Medium (SC), High (SD)	EPR–RN	EPR	EPR (SC), RN (SD)
	High income	High (SC), Medium (SD)		RN	RN (SC), EPR (SD)

Each row pair describes one session type (6 in total). Within a session, each P1 faces Observers (P2) drawn both from their own tier (socially close, SC) and from the other tier (socially distant, SD), so that across 24 rounds each participant interacts 12 times with SC and 12 times with SD counterparts. See Appendix Table 2.20 for the mapping between departments and categories.

Our decision to use these mechanisms is based on existing research showing that socioeconomic status is important to an individual's identity (Easterbrook et al., 2019) and that subjective perceptions of social status enhance identity formation (Kraus et al., 2011). For instance, Manstead (2018) found that material living conditions influenced key aspects of social behaviour. On the other hand, there is evidence on the strong link between political preferences and cheating in politics, including among political parties (Inglehart, 2000) and governments (Van De Walle, 2008). Identification with political parties is a key decisive factor in elections (Lutz and Lauener, 2020) and they tend to uphold over time, rendering them a reliable measure (Dalton, 2016).

Objective social closeness

Objective closeness is implemented at the time of recruitment and is how experimental sessions were structured. We coded is as a binary variable for both treatments (socially close vs socially distant).

Treatment 1: Socioeconomic Status

We recruited participants residing in Paris, Strasbourg and several departments of the *Île-de-France* region (see Table 2.20). We used official departmental income statistics from the French *National Institute of Statistics and Economic Studies (INSEE)* to construct SES tiers. Departments were assigned to tiers based on INSEE's 2021 *niveau de vie annuel médian* (median disposable income per consumption unit), anchored to (i) the national median (1,993€/month) and (ii) the highest-income departments in our sample (Paris and Hauts-de-Seine, both over 2,477€/month) (INSEE Références, 2021):

- **Low SES**: departments with median income below 1,993€ and an unemployment rate above 10% (Seine-Saint-Denis).
- **Medium SES**: departments with median income between 1,993€ and 2,477€, and an unemployment rate around 7% (Essonne, Seine-et-Marne, Val-de-Marne, Val-d'Oise, and Strasbourg).

- **High SES**: departments with median income above 2,477€ and an unemployment rate of 6% or less (Paris and Hauts-de-Seine).

In each round of the DUTC, participants were reminded of the average annual income and unemployment rate for their own department and for their counterpart's department, together with their classification as low, medium or high-income. For both, we also displayed national reference values (average annual income and unemployment rate in France) as a common benchmark. Department names were not shown to avoid stereotype-based responses.

Treatment 2: Political party alignment

For the political treatment, objective closeness is based on the political affiliation of the deputy elected in each participant's constituency in the 2024 French legislative elections. Participants were recruited from Paris, Île-de-France and Nice to capture variation across the three main blocs that dominated these elections:

- ***Nouveau Front Populaire* (NFP)**: a broad left-wing alliance, elected in eastern Parisian constituencies;
- ***Ensemble pour la République* (EPR)**: the centre-right presidential alliance, elected in western Paris and the Yvelines;
- ***Rassemblement National* (RN)**: the far-right party and right-wing allies, elected in Nice.

The repeated on-screen information about one's own and the counterpart's SES tier or deputy's party is intended to make these identity cues salient at the moment of choice, thereby mimicking contexts where income background or partisan alignment are chronically accessible categories in social evaluation.

Subjective social closeness

Subjective closeness is measured *ex-post*, using participants' self-reported socioeconomic and political profiles in Task 1 and aligning them with the objective tiers of their counterpart's locality.

Treatment 1: Socioeconomic Status

For subjective SES, we construct an individual SES tier for each participant based on self-reported income in Task 1. When participants report no personal income (e.g. students, unemployed), we use household or partner income as a proxy; if neither is available, we classify them as Low SES. This procedure assigns each participant to one of three SES tiers: Low, Medium, or High.

For comparability with the counterpart's locality, these three tiers are then indexed onto a 0–2 scale: Low = 0, Medium = 1, High = 2. Subjective SES closeness is defined by comparing the participant's own tier to the SES tier of the counterpart's department of residence (also coded 0–2). We construct two measures:

- **Binary subjective distance (SES)**: equal to 0 (close) if the participant's tier matches the counterpart's locality tier and 1 (distant) otherwise.
- **Ordered subjective distance (SES)**: step-distance given by the differences between tiers,

$$\text{SES subj distance} = \left|\texttt{participant SES tier} - \texttt{counterpart SES tier}\right| \in \{0, 1, 2\},$$

0 indicates the same tier; 1 indicates adjacent tiers or "short-distance" (e.g. Low–Medium or Medium–High); and 2 indicates a two-step separation or "long-distance" (Low–High).

Treatment 2: Political alignment

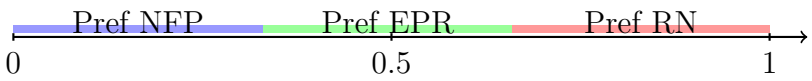
For subjective political closeness, we start from participants' 0–10 agreement ratings with NFP, EPR and RN in Task 1 (i.e., their *raw* political preferences, see Figure 2.13). We then construct a normalised political score in two steps. Let r_{NFP}, r_{EPR}, r_{RN} denote the three 0–10 sliders of raw preferences. We first convert them into party weights that sum to one:

$$w_{\text{NFP}} = \frac{r_{\text{NFP}}}{r_{\text{NFP}} + r_{\text{EPR}} + r_{\text{RN}}}, \quad w_{\text{EPR}} = \frac{r_{\text{EPR}}}{r_{\text{NFP}} + r_{\text{EPR}} + r_{\text{RN}}}, \quad w_{\text{RN}} = \frac{r_{\text{RN}}}{r_{\text{NFP}} + r_{\text{EPR}} + r_{\text{RN}}}.$$

We then place the three blocs on an ideological axis with NFP at 0, EPR at 0.5 and RN at 1, following a left-centre-right rationale in traditional politics, and compute a single political score

$$\text{pol\_score}_i = 0 \times w_{\text{NFP},i} + 0.5 \times w_{\text{EPR},i} + 1 \times w_{\text{RN},i},$$

which lies in $[0, 1]$. Intuitively, this is the expected ideological position of individual i given the distribution of their party support; values near 0, 0.5 and 1 correspond to predominant alignment with NFP, EPR and RN respectively:



For comparability with the observer's locality tier (0 = NFP, 1 = EPR, 2 = RN), we discretise pol_score_i into 0–2 tiers in two ways:

- **Relative tiers**: an asymmetric split that mirrors the empirical distribution of political scores in our sample,

$$\text{Relative} = \begin{cases} 0 & \text{if pol\_score}_i < 0.23, \\ 1 & \text{if } 0.23 \leq \text{pol\_score}_i < 0.50, \\ 2 & \text{if pol\_score}_i \geq 0.50. \end{cases}$$

  This yields three groups interpreted as left-leaning (0), centre-right (1) and far-right (2).

- **Absolute tiers**: a symmetric discretisation using terciles on the $[0, 1]$ spectrum,

$$\text{Absolute} = \begin{cases} 0 & \text{if pol\_score}_i < 0.25, \\ 1 & \text{if } 0.25 \leq \text{pol\_score}_i < 0.75, \\ 2 & \text{if pol\_score}_i \geq 0.75. \end{cases}$$

In simple terms, we first place each participant on a [0,1] NFP–EPR–RN spectrum, then convert that into three-step tiers in two ways (relative and absolute) and finally measure how close each tier is to the observer's local political tier: same, one step away, or two steps away, with an additional continuous gap in the absolute specification.

Estimation sample, roles and dependence

Unless otherwise noted, regressions linking social distance to cheating focus on Die-rollers (P1), since only P1's reports directly measure misreporting in the DUTC. To account for within-individual and within-session dependence, standard errors are clustered at the participant level (and, in robustness checks, at the session level). Regressions using social distance and closeness variables are restricted to treatment sessions, where objective and subjective SES or political tiers are defined; the control group—where no SES or political manipulation is implemented—is used as a descriptive benchmark but is not directly comparable on the distance scales.

Experimental procedures

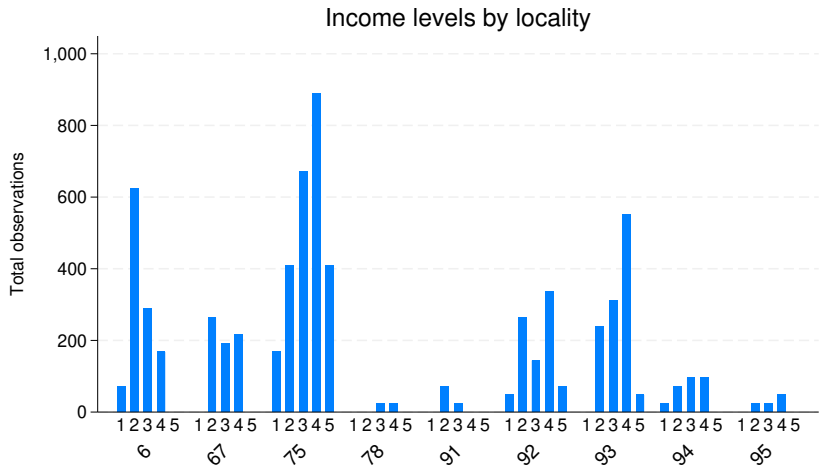
The experiment was conducted online in December of 2024. We ran 6 sessions with 288 voluntary participants (53% female) with an average age of 30 years (st. dev. 11). Participants were mostly students (41%) and employed individuals (37%). For detailed information on scholarity levels and occupation of participants, refer to Appendix Tables 2.21 and 2.22. Average earnings in the experiment were 6.87 euros (st. dev. 5.84), including a 4€ show up fee.

2.4 Descriptive statistics

2.4.1 Socioeconomic and political data in the sample population

Figure 2.11 displays the distribution of participants' income levels across localities of residence, coded as follows: 1 = monthly income below €1,000; 2 = €1,000 to 2,000; 3 = €2,000 to 4,000; 4 = €4,000 to 6,000; and 5 = above €6,000. Each cluster of bars represents the distribution of income levels within a locality, identified by postal code: 6 (Nice), 67 (Strasbourg), 75 (Paris), 78 (Yvelines), 91 (Essonne), 92 (Hauts-de-Seine), 93 (Seine-Saint-Denis), 94 (Val-de-Marne and Seine-et-Marne, due to the lower number of participants) and 95 (Val-d'Oise). The graph shows that our sample was relatively diverse, with a high concentration of observations in Paris and substantial representation from other Île-de-France departments such as Hauts-de-Seine and Seine-Saint-Denis, which were all used in the socioeconomic treatment.

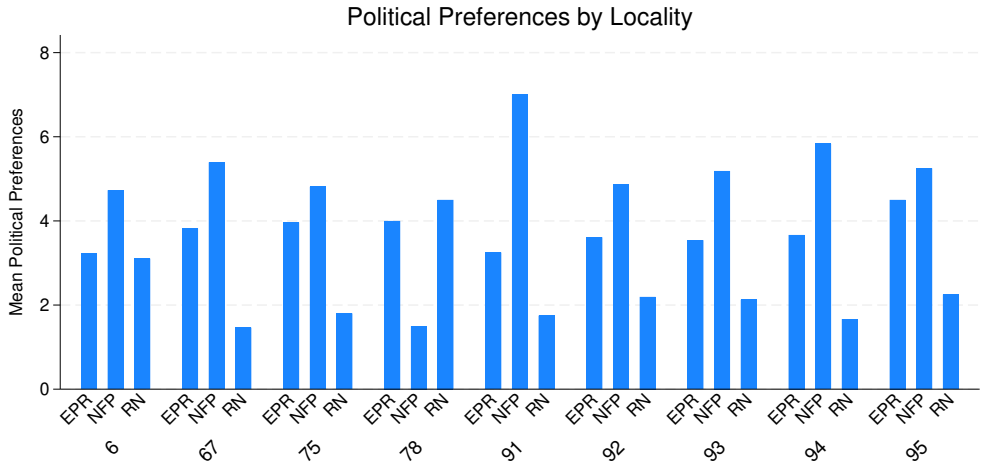
Figure 2.11: Distribution of income levels by locality



Sample: $N = 6{,}912$ observations. By postcode - Paris: 2,544; Nice & Seine-Saint-Denis: 1,152 each; Hauts-de-Seine: 864; Strasbourg: 672; Marne: 288; Essonne & Val-d'Oise: 96 each; Yvelines: 48. Income tiers: 1: 312 (4.51%), 2: 1,968 (28.47%), 3: 1,776 (25.69%), 4: 2,328 (33.68%), 5: 528 (7.64%).

Figure 2.12 presents average political preferences across localities. Participants rated their support (on a 0–10 scale) for the three political forces in the 2024 French legislative elections: the left-wing alliance *Nouveau Front Populaire* (NFP), the centrist presidential alliance *Ensemble pour la République* (EPR) and the far-right party *Rassemblement National* (RN). Our sample represented substantial variations across localities, with stronger average support for EPR in departments like Yvelines and Paris and high NFP preferences in Paris. Preferences for RN were the lowest in all localities.

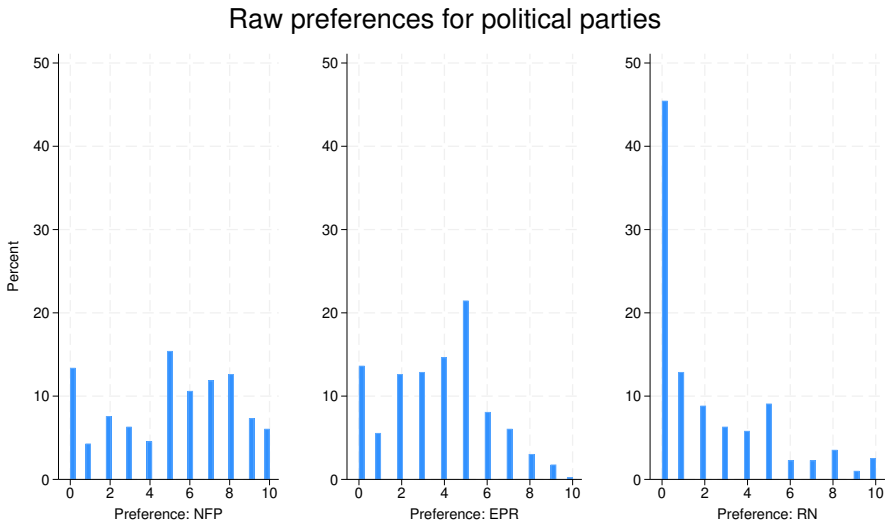
Figure 2.12: Political preferences by locality



The graph shows the distribution of average political preferences by department. Taking the $N = 6{,}912$ observations in the sample as benchmark, we report mean preferences for the departments with 100+observartions: Paris (2,544 observations) with mean preferences of NFP = 4.81, EPR = 3.96, and RN = 1.80; Seine-Saint-Denis (1,152 obs) (NFP = 5.18, EPR = 3.54, RN = 2.12), Nice (1,152 obs) (NFP = 4.72, EPR = 3.22, RN = 3.10); Hauts-de-Seine (864 obs) (NFP = 4.86, EPR = 3.61, RN = 2.19); Strasbourg (672 obs) (NFP = 5.39, EPR = 3.82, RN = 1.46); and Marne (288 obs) (NFP = 5.83, EPR = 3.66, RN = 1.66).

Figure 2.13 shows raw preferences for each political party across the full sample, taken directly from Task 1 on a 0–10 scale. The distribution indicated that participants were more evenly distributed in their preferences for NFP and EPR, whereas a large proportion reported near-zero preferences for RN.

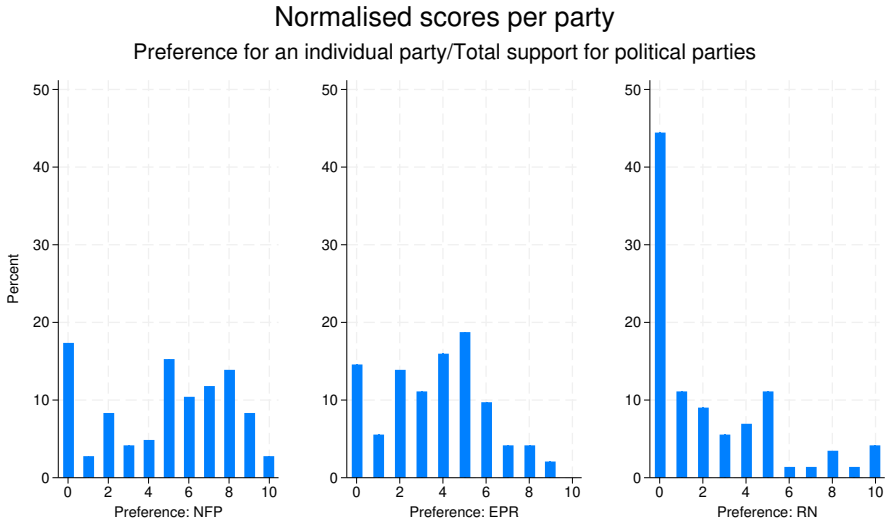
Figure 2.13: Raw political preferences



The graph shows raw preferences for each party in the entire sample (N = 6,912): NFP: Mean = 4.98, Std. Dev. = 3.09; EPR: Mean = 3.70, Std. Dev. = 2.34, RN: Mean = 2.11, Std. Dev. = 2.79.

Figure 2.14 shows the normalised political score for each party $j \in \text{NFP}, \text{EPR}, \text{RN}$, calculated as each individual's preference divided by the total support for all three parties. This measure captures subjective alignment on a continuous scale from 0 to 1. The distributions indicate that support for EPR and NFP is relatively balanced, whereas RN is highly skewed, with nearly half of participants assigning it a score of zero, reflecting limited alignment with this party in the sample.

Figure 2.14



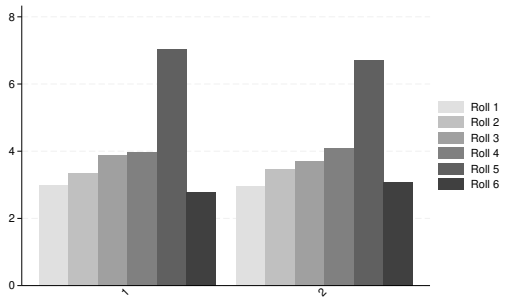
2.4.2 Propensity to cheat in the DUTC

Figure 2.15 reports the frequency distribution of the six die-roll outcomes in DUTC. As predicted, we observe that participants over-reported the payoff-maximising outcome (5: mean = 7.05 in T1, 6.70 in T2) while other outcomes were reported closer to each other (outcomes 1–6: means ranging between 2.78–4.08). In both instances, participants deviated from the uniform statistically distribution (mean = 2.5), obtained through the following formula:

$$Mean = \frac{1}{6} \times (1 + 2 + 3 + 4 + 5 + 0) = \frac{15}{6} = 2.5$$

The average reported payoff across the entire sample was 3.71 (SD = 1.59), indicating that on average participants inflated their die-roll reports. When disaggregating results by treatment, we observed virtually no differences: T1: mean = 3.71 and T2: mean = 3.72

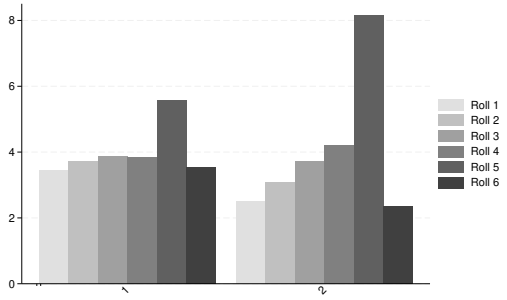
Figure 2.15: Frequency of reported rolls in the DUTC by treatment



In the graph, 1 indicates rolls reported by Die-rollers (P1) (N = 3,456 obs), 2 indicates rolls reported by Die-rollers (P2) (N = 3,456 obs)

Figure 2.16 shows the distribution of reported payoffs by player role. Results show that, while the trends on over-reporting persist, the likelihood to cheat was much larger for observers (mean = 3.22) than for die-rollers (mean = 2.74). Notably, P2 reported rolling a payoff of 5 an average of 8.16 times, well above the statistical expectation and the corresponding averages for P1 (5.58). Moreover, differences in deviations from expected payoff by role were significant at the 1% level (p-value = 0.003), indicating that the role of observer largely influenced participants' propensity to cheat. The regression, however, explains a small proportion of the variance in the dependent variable ($R^2 = 0.046$), which we later disseminate more in detail.

Figure 2.16: Frequency of rolls in the DUTC by player type



In the graph, 1 indicates Treatment 1 (Socioeconomic) (N = 3,456), 2 indicates Treatment 2 (Political)(N = 3,456)

Figure 2.17 displays the distribution of the normalised deviation measure. For each Die-roller, we first compute their average reported die value across the 24 rounds and compare it to the 2.5 benchmark. We then express this difference as a fraction of 2.5, so that a value of zero corresponds to truthful reporting on average, while a value of one would correspond to maximal profitable misreporting (only reporting 5). The right-skewed distribution suggests that while relatively few participants are perfectly truthful, most report above the expected outcome and repeatedly inflate their die-roll reports; the mean value of 0.573 (SD = 0.168) confirms moderate deviations from the truthful benchmark, in line with previous DUTC studies (Alfonso-Costillo et al., 2022; Fischbacher and Föllmi-Heusi, 2013; Siniver et al., 2022; Tobol et al., 2020).

Figure 2.17: Normalised distribution of cheating

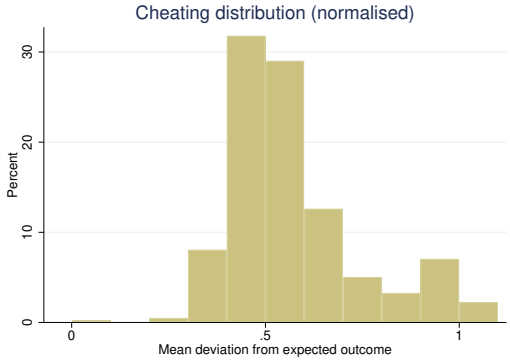
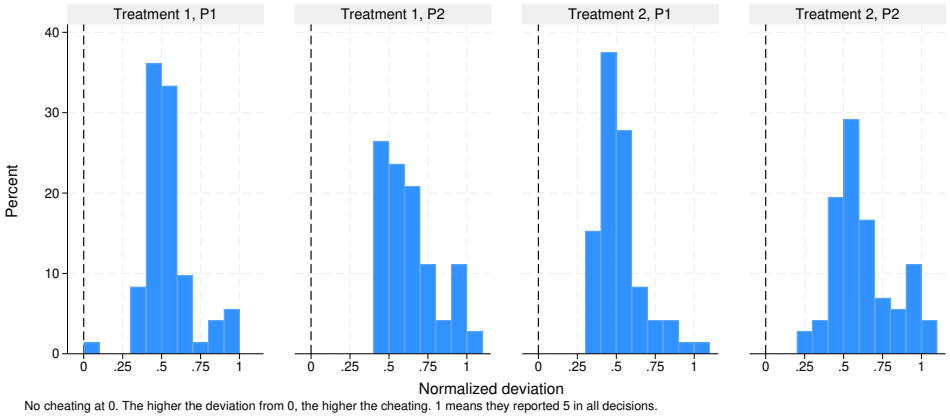


Figure 2.18 disaggregates the normalised deviation measure by treatment and player role. The figure suggests that the treatment manipulation altered participants' propensity to cheat and that this effect differs sharply by role: P1 participants display relatively moderate deviations clustered around lower values of the scale, whereas P2 participants exhibit substantially higher deviations. While some variation is visible across the treatments, overall cheating levels differ more between roles.

Figure 2.18: Cheating by treatment and player type



2.5 Results

Result 1. The association between social closeness and increased cheating

Result 1a. Cheating across objective social distance

We test **H1:** Social closeness is associated with higher propensity to cheat in outcome reports, in a unified regression frameworks, splitting analyses by objective and subjective measures of social closeness. We begin by examining whether cheating is higher when participants are paired with an objectively close counterpart (same SES or same political affiliation at the department level) using the dependent variable:

- **Cheating:** `gain`: reported payoff per round in the DUTC.

We use two distance regressors and report complementary within-participant checks:

- **Objective social distance (binary):** `binary_obj`, 0 if participants' departments share SES level/political affiliation, 1 if not.
- **Objective social distance (ordered):** `ternary_obj`, 0 if participants are socially close, 1 if SES or political affiliation in their departments is short distanced (Low–Medium or Medium–High SES; NFP–EPR or EPR–RN) and 2 if SES or political affiliation in their departments is short distanced (Low-High SES; NFP-RN).

A two-sample t-test indicates slightly higher reported gains when participants were socially close (mean = 2.781) versus socially distant (mean = 2.704) (diff. = 0.077, $t = 1.29$, two-sided $p = 0.197$). Regression results in the pooled model (T1+T2), with standard errors clustered by participant, show no significant effects: `binary_distant_obj` = -0.077, $p = 0.141$) and `ternary_distant_obj` = -0.036, $p = 0.392$).

When splitting analyses by treatment, results further show weak-to-marginal patterns. In the socioeconomic treatment (T1) the raw means show a modest difference between close and distant pairs (mean diff. = 0.160; $p = 0.059$). Regression results with participant-clustered errors shows that neither `binary_obj` (-0.160, $p = 0.029$), nor `ternary_obj` (-0.099, $p = 0.104$) are significant. By contrast, in the political treatment (T2) both the means and clustered regressions are essentially null (mean diff. $= -0.006$, $p = 0.945$) and both (`binary_obj` = 0.006, $p = 0.938$) `ternary_obj`= 0.019, $p = 0.742$). Overall, evidence that objective closeness raises reported gains is concentrated in the T1 binary specification.

By treatment, T1 shows borderline reductions with distance (`binary_obj`: -0.160, $p = 0.058$; `ternary_obj`: -0.099, $p = 0.084$), (all $p > 0.28$).

When restricting analyses to P1 observations, close and distant pairs again display very similar gains (pooled P1 mean diff. = 0.077, two-sided $p = 0.197$) and regression results further confirm previously observed patterns (`binary_obj` = -0.077, $p = 0.141$; `ternary_obj` = -0.036, $p = 0.392$). Splitting by treatment, T1 again shows the strongest pattern: `binary_distant_obj`

= -0.160, $p = 0.029$, while the ordered specification remains non-significant (-0.099 $p = 0.104$), whereas all T2 specifications are non-significant ($p > 0.28$).

In sum, after accounting for within-participant clustering, objective closeness is associated with higher reported gains mainly in the socioeconomic treatment (binary encoding); pooled and political-treatment estimates are small and not robustly significant.

Table 2.14: Result 1a: Objective social distance associated with higher `gain`

	Pooled (T1+T2)		T1 (SES)		T2 (Political)	
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
`binary_obj`	-0.077		-0.160**		0.006	
	(0.052)		(0.072)		(0.075)	
`ternary_obj`		-0.036		-0.100		0.019
		(0.042)		(0.060)		(0.058)
Female (=1)	0.084	0.085	0.112*	0.109*	0.085	0.084
	(0.113)	(0.113)	(0.152)	(0.153)	(0.162)	(0.162)
Age	-0.008**	-0.008**	0.002	0.002	-0.012***	-0.012***
	(0.004)	(0.004)	(0.009)	(0.009)	(0.004)	(0.004)
Occupation	0.028	0.029	0.088*	0.090*	-0.012	-0.011
	(0.035)	(0.035)	(0.047)	(0.047)	(0.051)	(0.052)
Scholarity	-0.027	-0.026	0.005	0.006	-0.044	-0.044
	(0.046)	(0.046)	(0.062)	(0.062)	(0.070)	(0.070)
Constant	2.956***	2.939***	2.342***	2.317***	3.230***	3.222***
	(0.349)	(0.354)	(0.553)	(0.553)	(0.463)	(0.470)
Observations	3,456	3,456	1,728	1,728	1,728	1,728
R-squared	0.005	0.005	0.009	0.009	0.009	0.009

Robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Result 1b. Cheating across subjective social distance

We now examine whether cheating varies with the measures of subjective closeness. Our dependent variable remains (`gain`) and we use the following regressors in a unified framework:

- **Subjective social distance (binary):** `binary_subj`, 0 if the participant's individualised SES/political preference matches the SES/political affiliation in the counterpart's locality, 1 otherwise.
- **Subjective social distance (ordered):** `ternary_subj`, absolute tier gap between the participant's own tier and the counterpart's locality tier.
- **Additional binary/ternary measures T2:** `binary`$_s ubj_A and$

Results in the pooled model across T1+T2 show that `gain` means are virtually identical across subjectively close (2.766) and distant pairs (2.730) (diff. 0.036, $t = 0.58$, $p = 0.565$). Similarly, regression analysis confirms null effects (`binary_subj` = -0.044, $p = 0.560$; `ternary_subj` = -0.005, $p = 0.912$).

These results repeat when splitting analyses by treatment: in the socioeconomic treatment, close and distant means are very similar (diff.0.013, $p = 0.885$), while regression results are small and non-significant (`binary_subj` = -0.013, $p = 0.907$; `ternary_subj` = -0.004, $p = 0.950$); in the political treatment, results are likewise small and not significant (`binary_subj` = -0.086, $p = 0.420$; `ternary_subj` = -0.010, $p = 0.878$).

When restricting analyses to P1, results yield the same patterns. Pooled difference is small (diff. $= -0.013$, $p = 0.885$) and regressions are non-significant (`binary_subj` = -0.013, $p = 0.907$; `ternary_subj` = -0.004, $p = 0.950$. Coefficient in T1 remain near zero and not significant ($p = -0.086$; $p = 0.420$), and all T2 specifications are null ($p = 0.420$; $p = 0.878$).

Overall, across pooled, treatment-specific and reporter-only analyses, subjective social closeness does not increase cheating. Point estimates are near zero and consistently non-significant across binary, ordered, and continuous encodings.

Table 2.15: Result 1b: Subjective social distance associated with higher `gain`

	Pooled (T1+T2)		T1 (SES)		T2 (Political)			
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
`binary_subj`	-0.045		-0.013		-0.086			
	(0.076)		(0.112)		(0.106)			
`ternary_subj`		-0.006		-0.005		-0.010		
		(0.050)		(0.075)		(0.066)		
`binary_subj_A`							-0.017	
							(0.105)	
`ternary_subj_A`								-0.045
								(0.069)
Female (=1)	0.087	0.085	0.113	0.113	0.092	0.085	0.086	0.086
	(0.113)	(0.112)	(0.151)	(0.150)	(0.163)	(0.163)	(0.161)	(0.162)
Age	-0.008**	-0.008**	0.002	0.002	-0.012***	-0.012***	-0.012***	-0.012***
	(0.004)	(0.004)	(0.009)	(0.009)	(0.004)	(0.004)	(0.004)	(0.004)
Occupation	0.029	0.029	0.088*	0.088*	-0.011	-0.011	-0.011	-0.009
	(0.035)	(0.035)	(0.048)	(0.048)	(0.052)	(0.051)	(0.052)	(0.052)
Scholarity	-0.028	-0.027	0.005	0.005	-0.048	-0.044	-0.044	-0.043
	(0.047)	(0.046)	(0.062)	(0.063)	(0.070)	(0.071)	(0.070)	(0.070)
Constant	2.943***	2.920***	2.269***	2.264***	3.294***	3.242***	3.242***	3.253***
	(0.356)	(0.352)	(0.559)	(0.553)	(0.463)	(0.473)	(0.463)	(0.460)
Observations	3,456	3,456	1,728	1,728	1,728	1,728	1,728	1,728
R-squared	0.005	0.005	0.007	0.007	0.009	0.009	0.009	0.009

Robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Manipulation check on social closeness

To verify that our SES and political-alignment manipulations elicited feelings of relatedness, we examined participants' responses on the 1-to-7 IOS scale (`IOS_same`: closeness to counterparts from the same locality; `IOS_diff`: from different localities). We computed a unified variable `ios_score`, by computing the difference between these two measures: `playercloseness_same - playercloseness_diff`, positive values indicate that a participant feels closer to same-locality

counterparts, negative values indicate closer to different-locality counterparts, and zero indicates equally perceived closeness.

Participants classified as socially close reported higher perceptions of closeness to same-department counterparts (mean `ios_score` = 1.119) than participants classified as socially distant (mean `ios_score` = 0.781); the mean difference is 0.338 points, which is not statistically significant ($p = 0.161$). Regression analyses confirmed these effects: social closeness was associated with an increase in `ios_score` of $\beta = 0.338$ ($p = 0.157$) relative to distant participants. These results indicate that the SES and political-alignment cues were associated in the expected direction with reported relatedness, although the difference is not statistically significant. For raw IOS scores, refer to Figure 2.19 in the Appendix.

Table 2.16: Manipulation check: IOS scores by objective social distance

Distance	*N*	Mean IOS_same	Mean IOS_diff	SD (IOS_same / IOS_diff)
Distant (0)	144	3.570	2.789	1.746 / 1.581
Close (1)	144	3.556	2.438	1.808 / 1.457
Difference	–	-0.014	-0.351	–

IOS = Inclusion of Other in the Self scale. Higher scores indicate greater perceived closeness.

Result 2. Observation by a socially close counterpart and cheating

We now test **H2:** Cheating in outcome reports is higher when participants are observed by socially close counterparts. Therefore, we focus on Player 1 (the die-roller) by comparing behaviour across two observation blocks: 12 rounds with socially close observers and 12 rounds with socially distant observers. We use block-level distance regressors. Since each participant plays one close and one distant observation block, coding the distant block as 1 (close = 0) estimates the within-participant change in average gain from being observed by close to distant counterparts:

- **Objective social distance (binary):** `binary_obj_block`, 0 in the close block, 1 in the distant block; a negative coefficient means lower cheating when the pair is distant.
- **Objective social distance (ordered):** `ternary_obj_block`, 0 if close, 1 if short distance (adjacent tiers, e.g.Low–Medium SES or NFP–EPR) and 2 if long distance (two-step tiers, e.g. Low–High SES or NFP–RN); a negative value implies lower gains as distance increases (i.e., relatively more cheating when close).
- **Subjective relative distance:** `obs_subj_bin`, a binary match between the participant's own tier and the observer's locality tier (0=close); and `obs_subj_ord`, the ordered step gap (0= same tier; 1=adjacent; 2=two-step).
- **Subjective absolute distance (T2 only):** uses the participant's political score (NFP=0, EPR=0.5, RN=1) mapped to the locality's 0–2 tier. We create three variables:
  - `subj_abs_distant_bin`: binary block code (0 = close, 1 = distant).

– `subj_abs_distance_steps`: ordered step distance (0/1/2).

– `subj_abs_distance_gap`: continuous absolute distance $\left|\, 2 \cdot \text{political score} - \text{locality tier} \,\right|$.

Mean gains in the pooled model (T1+T2) are 2.781 under close observation and 2.704 under distant observation (diff. =0.076, $t = 1.289$, $p = 0.197$). In the regression model, the effects of objective closeness are small and not significant: `obs_obj_bin` = 0.076, $p = 0.716$) and `obs_obj_ord` = 0.076, $p = 0.603$). This pattern repeates for subjective closeness: `obs_subj_bin` = 0.018, $p = 0.937$; `obs_subj_ord` = 0.187, $p = 0.287$. Thus, in the pooled model, there is no evidence that observation by close counterparts raises cheating, if anything, the non-significant results lean toward slightly higher gains under distant observation.

When disaggregating results by treatment, effects in the socioeconomic treatment (T1) are positive and statistically significant, indicating more cheating when the observer is socially distant (`obs_obj_bin` = 0.583, $p = 0.041$; `obs_obj_ord` = 0.387, $p = 0.043$). Subjective relative measures remain small and non-significant (`obs_subj_bin` = -0.144, $p = 0.642$; `obs_subj_ord` = -0.140, $p = 0.640$).

Table 2.17: Result 2: Observation by socially close P2 (pooled)

	Objective distance		Subjective distance	
Variables	(1)	(2)	(3)	(4)
`obs_obj_bin`	0.076			
	(0.210)			
`obs_obj_ord`		0.076		
		(0.145)		
`obs_subj_bin`			0.018	
			(0.223)	
`obs_subj_ord`				0.187
				(0.175)
Gender (Female=1)	0.099	0.099	0.098	0.167
	(0.202)	(0.201)	(0.202)	(0.244)
Age	0.001	0.001	0.001	0.005
	(0.010)	(0.010)	(0.010)	(0.010)
Occupation	0.020	0.019	0.019	0.020
	(0.067)	(0.067)	(0.067)	(0.080)
Scholarity	0.125	0.124	0.125	0.205*
	(0.092)	(0.092)	(0.093)	(0.109)
Constant	1.684**	1.677**	1.712**	0.849
	(0.730)	(0.725)	(0.729)	(0.802)
Observations	288	288	288	216
R-squared	0.007	0.008	0.007	0.024

Robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

As for results in the political treatment (T2), objective closeness effects are negative but non-significant (`obs_obj_bin` = -0.431, $p = 0.157$; `obs_obj_ord` = -0.250, $p = 0.250$). Subjective relative closeness are near zero (`obs_subj_bin` = 0.174, $p = 0.596$) and marginally positive for the ordered relative measure (`obs_subj_ord` = 0.372, $p = 0.083$). Under the absolute scaling,

the binary and continuous measures remain non-significant (`subj_abs_distant_bin` = 0.357, $p = 0.275$; `subj_abs_distance_gap` = 0.332, $p = 0.220$), while the ordered absolute distance is positive and statistically significant (`subj_abs_distance_steps` = 0.811, $p = 0.002$), implying higher cheating as absolute distance increases.

In sum, the pooled analyses show no reliable evidence that being observed by a socially close counterpart increases cheating. In treatment-specific splits, results in T1 show the opposite pattern, with significantly higher gains when observed by a distant counterpart; while in T2 the only robust effect appears for the ordered absolute subjective distance, again pointing to more cheating with greater social distance. Overall, effect sizes are modest and not consistently robust across measures, offering little support for H2 and suggesting weak evidence of the reverse in some specifications.

Table 2.18: Result 2: Observation by socially close P2 (Treatment 1)

	Objective distance		Subjective distance	
Variables	(1)	(2)	(3)	(4)
obs_obj_bin	0.583**			
	(0.282)			
obs_obj_ord		0.387**		
		(0.190)		
obs_subj_bin			-0.144	
			(0.309)	
obs_subj_ord				-0.140
				(0.299)
Gender (Female=1)	-0.202	-0.188	-0.195	0.179
	(0.261)	(0.261)	(0.265)	(0.402)
Age	-0.022	-0.022	-0.021	-0.010
	(0.014)	(0.015)	(0.014)	(0.018)
Occupation	-0.002	-0.010	0.002	0.073
	(0.092)	(0.093)	(0.092)	(0.141)
Scholarity	0.025	0.019	0.025	0.155
	(0.116)	(0.117)	(0.123)	(0.184)
Constant	2.961***	3.041***	3.322***	1.668
	(1.095)	(1.098)	(1.129)	(1.712)
Observations	144	144	144	72
R-squared	0.046	0.046	0.018	0.027

Robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2.19: Result 2: Observation by socially close P2 (Treatment 2)

	Objective distance		Subjective distance (relative)		Subjective distance (absolute)		
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
obs_obj_bin	-0.431						
	(0.302)						
obs_obj_ord		-0.250					
		(0.217)					
obs_subj_bin			0.174				
			(0.327)				
obs_subj_ord				0.372*			
				(0.213)			
obs_abs_bin					0.357		
					(0.326)		
obs_abs_ord						0.811***	
						(0.257)	
obs_abs_gap							0.332
							(0.270)
Gender (Female=1)	0.378	0.390	0.362	0.335	0.350	0.222	0.407
	(0.304)	(0.306)	(0.311)	(0.309)	(0.311)	(0.392)	(0.312)
Age	0.009	0.009	0.009	0.006	0.009	0.014	0.005
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.015)	(0.012)
Occupation	-0.041	-0.046	-0.043	-0.065	-0.049	-0.188	-0.045
	(0.100)	(0.100)	(0.101)	(0.101)	(0.102)	(0.121)	(0.103)
Scholarity	0.206	0.210	0.214	0.219	0.211	0.577***	0.193
	(0.147)	(0.146)	(0.146)	(0.144)	(0.142)	(0.188)	(0.143)
Constant	1.209	1.146	0.871	0.844	0.817	-0.847	0.854
	(1.008)	(1.005)	(1.009)	(0.949)	(0.953)	(1.101)	(0.939)
Observations	144	144	144	144	144	72	142
R-squared	0.044	0.040	0.032	0.051	0.038	0.220	0.039

Robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

2.6 Concluding Remarks

In this paper we sought to test whether social closeness fosters dishonest reporting in a repeated online DUTC, and whether being observed by a socially close counterpart amplifies this behaviour. Implementing two real-world dimensions of closeness (socioeconomic status and political alignment), and constructing binary and scaled indicators, we intended to enlarge the scope of our analyses to fill up gaps we identified in the literature on social preferences, in-group favouritism and observability.

Across pooled and treatment-specific analyses, we find little evidence that social closeness systematically increases cheating. Similarly, objective closeness shows, at most, a small and statistically marginal pattern, which is sensitive to specification and does not uphold consistently within either the socioeconomic or political treatments. Moreover, subjective closeness did not render significant in testing our hypotheses, with mean differences nearing zero and small and non-significant regression coefficients. Our results hold both in pooled samples and when restricting attention to outcome reports, therefore, pointing to limited and context-dependent associations between social proximity and misreporting in this setting.

Our findings complement and nuance the broader literature. Classic DUTC studies show that misreporting responds to incentives, feedback, and strategic context (e.g., Benistant et al. (2021); Fischbacher and Föllmi-Heusi (2013); Siniver et al. (2022)), while research on social proximity documents that in-group ties can foster coordination, favouritism, and even collaborative dishonesty in some designs Gross and De Dreu (2021); Weisel and Shalvi (2015); **?**. In our real-world closeness manipulations, however, any motivation to cheat enticed by in-group preferences appears weak, as objective distance yields only fragile patterns and subjective closeness is consistently null. This suggests that many theorised channels previously explored in the literature, such as in-group justification, prosocial cheating or leniency towards similar others may require stronger cues than those present in our neutral-payoff, anonymous DUTC environment. This opens the door to the exploration of norm enforcement mechanisms or explicit group payoffs as potential cues to identify the expected outcomes.

Methodologically, the study offers a multi-scale operationalisation of social closeness and a clean observation design at the individual level, enabling a clear separation of between-subject differences from within-person behavioural changes. The small effects and lack of robustness, however, advise caution when attributing dishonest behaviour to everyday forms of social proximity, namely to shared SES tier or people's political colours. For applied contexts (work teams, organisations, personal interactions) our results imply that simply pairing similar individuals may not induce more cheating when payoffs are individualised, independent between participants and, more importantly, when monitoring is passive.

In conclusion, within a repeated DUTC where payoffs are individual and observation is passive, social closeness, whether objectively or subjectively defined, does not reliably increase cheating, nor does close observation meaningfully shift die-rollers' behaviour relative to distant observation. Any relationship between social proximity and dishonesty in this environment appears limited and contingent, underscoring the need for multi-method measurement and careful design when evaluating how social ties shape unethical behaviour.

Limitations and future directions. We identify several limitations to future work. First, external validity beyond the DUTC and our specific SES and politics framings remains to be established. For instance, designs with dependent payoffs, norm enforcement, or salient group goals could reveal stronger social channels. Second, richer subjective measures such as multi-item identity scales or *ex-ante* identification and recruitment of participants from specific socioeconomic and political backgrounds could bolster feelings of relatedness. Third, registering true rolled values, while diminishing cleanness in observations, might elicit a clearer picture of how social closeness and observation by socially close counterparts associate with increased cheating.

Policy implications. Our results show that homophily (same SES tier or political alignment) is not by itself a meaningful cheating risk under passive observation. Organisations should therefore prioritise monitoring and clear rules among individuals, transparent scoring and credible detection of transgression. The adoption of low-cost and general safeguards in the likes of anomaly feedback, occasional anonymised review or integrity prompts, might be key in raising honesty regardless of who observes whom. In sum, the data suggest that for everyday settings, institutions could make cheating harder and costlier through clarity, transparency and credible detection.

2.7 Appendix

2.7.1 Tables

Table 2.20: Observations by locality and treatment

Locality of residence	**Socioeconomic treatment**	**Political treatment**
Paris	288	2,232
-Eastern Paris	0	*1,152* (NFP)
-Western Paris	*288* (H)	*1,104* (EPR)
Essonne	96 (M)	0
Hauts-de-Seine	864 (H)	0
Seine-et-Marne/Val-de-Marne	288 (M)	0
Seine-Saint-Denis	1,152 (L)	0
Val d'Oise	96 (M)	0
Yvelines	0	48 (EPR)
Nice	0	1,152 (RN)
Strasbourg	672 (M)	0
Other	0	0
Total	**3,456**	**3,456**

The table registers the total observations based on participants' localities of residence. Paris was split between east and west for T2 owing to marked differences in the electoral results from both zones (east largely voted for NFP, west largely voted for EPR), while for T1 the western districts with higher income were used for the high income level (H). The eastern Paris districts used in our sample were the 3rd, 4th, 11th, 12th, 13th, 14th, 18th, 19th and 20th *arrondissements*, while the western Paris districts were the 5th, 7th, 8th, 9th, and 16th *arrondissements*. Participants who came from localities other than the listed above were placed in the "Other" category.

Table 2.21: Demographics on scholarity

Diploma	session 1	session 2	session 3	session 4	session 5	session 6	Total
PhD (8 years)	48	72	48	0	0	24	192
Master (5 years)	456	528	624	528	552	672	3,360
Bachelor (3 years)	432	264	240	456	312	288	1,992
Technical degree (2 years)	96	144	192	72	168	24	696
High school	96	120	48	72	96	144	576
Secondary education	24	24	0	0	0	0	48
No diploma	0	0	0	24	24	0	48
Total	1,152	1,152	1,152	1,152	1,152	1,152	6,912

The total numbers refer to observations, to divide between the 48 participants in each session.

Table 2.22: Demographics on occupation

Occupation	session 1	session 2	session 3	session 4	session 5	session 6	Total
Employed	480	480	312	408	528	408	2,616
Independent	120	144	96	72	48	120	600
Student	384	432	576	552	456	456	2,856
Unemployed	0	48	48	24	24	24	168
Pensioner	144	24	120	48	96	144	576
Other	24	24	0	48	0	0	96
Total	1,152	1,152	1,152	1,152	1,152	1,152	6,912

Figure 2.19: Distribution of mean IOS scores by treatment

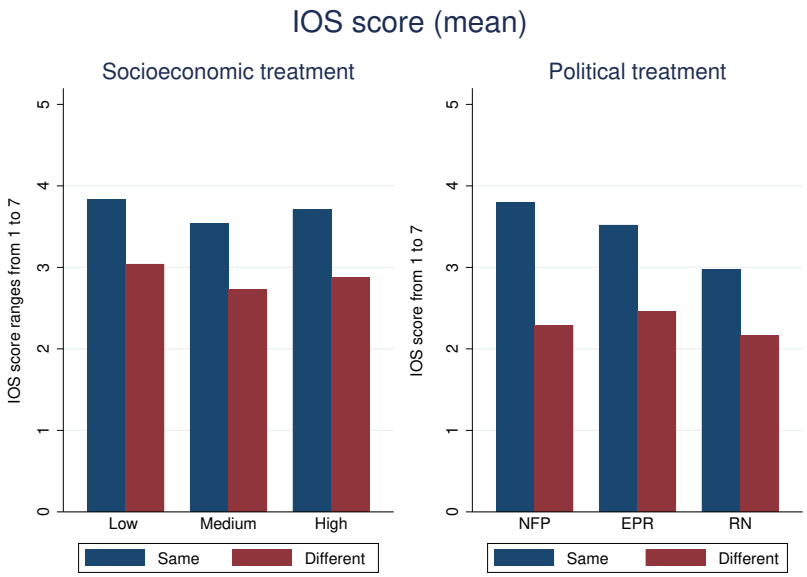


Figure 2.20: Originally registered protocol

ASPREDICTED

Wharton CREDIBILITY LAB
University of Pennsylvania

How Social Closeness Fuels Dishonesty (#200517)

Author(s)
This pre-registration is currently anonymous to enable blind peer-review.
It has 3 authors.

Pre-registered on:
2024/11/21 01:41 (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
This study investigates whether being observed by a socially close individual influences the propensity to misreport outcomes in a Die-under-the-cup (DUTC) game. Specifically, it examines whether participants observed by a socially close individual are more likely to misreport their die roll outcomes to increase their gains compared to those observed by a socially distant individual. The project answers two research questions:

Does observation by a socially close individual increase the propensity to cheat in the DUTC game?
Do perceptions of social closeness mediate the propensity to cheat when observed by a socially close vs a socially distant individual?

3) Describe the key dependent variable(s) specifying how they will be measured.
The dependent variable is the propensity distribution for cheating in the DUTC, measured by deviations to statistically expected outcomes of a fair die roll (17% for each side).

The study has two independent variables:
Social Closeness (binary): observed by someone from the same locality (Socially Close, SC) or from a different locality (Socially Distant, SD).
Perceptions of Social Closeness (continuous): perception of closeness based on locality, measured using scores in the Inclusion of the Other in the Self scale (IOS).
Additionally, we use the Moral Identity Scale (MIS) to control for internalisation of moral traits.

4) How many and which conditions will participants be assigned to?
Control group: Participants roll a virtual die and register the outcome with no observation. We register the average over all the rolls made in 24 rounds

Two treatment Conditions:
Socioeconomic treatment: socioeconomic data about own and counterpart's locality of residence. Three cohorts: high, medium and low-income localities.
Political preference treatment: data on political preferences in own and counterpart's locality of residence. Three cohorts: left-wing, centre-right and far-right party.

Participants are assigned to one of two roles:
Player 1 (P1): Rolls the virtual die in private and reports the outcome to P2.
Player 2 (P2): Observes the roll outcome reported by P1 and registers a single final outcome.
We register the distribution of 144 die rolls corresponding to 12 rolls of 12 P1 matched with 12 P2.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
For our first question, we compute distributions of expected die-roll frequencies (cheating) segmented by role (P1, P2) and treatment (control, socioeconomic, political). We conduct a chi-square test to measure deviations in propensity to cheat from the expected uniform distribution when P1 participants are observed by a socially close (SC) or socially distant (SD) P2:
H0: The propensity to cheat is independent of whether participants are observed by a SC or SD.
HA: The propensity to cheat depends on whether participants are observed by a SC or SD.

For our second question, we use causal mediation analysis modelling social closeness as a binary independent variable (SC vs SD), perceptions of closeness as the mediator (continuous variable IOS scores) and propensity to cheat as the dependent variable. We assess the following casual relations:
Direct Effect (DE): The impact of social closeness on the propensity to cheat does not pass through perceptions of social closeness.
Indirect Effect (IE): The impact of social closeness on the propensity to cheat is mediated by perceptions of social closeness.
Total Effect (TE): Combined effect of social closeness on cheating (TE = DE + IE).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
NA

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
We recruit 348 subjects, 144 per treatment and 60 in the control. Our sample size consists of 6,912 observations – 3,456 observed by SC and 3,456 by SD participants, plus 1,440 observations in the control group.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

We run a two-way ANOVA test if cheating propensity differs within the treatments (e.g. high-income vs low-income localities, politically aligned localities), as well as an additional chi-square to test if P2 participants (observers) are more likely to adjust outcomes when observing SC versus SD individuals.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4):1115–1153.

Accinelli, E. and Carrera, E. J. (2012). Corruption driven by imitative behavior. *Economics Letters*, 117(1):84–87.

Acedo-Carmona, C. and Gomila, A. (2014). Personal trust increases cooperation beyond general trust. *PLoS ONE*, 9(8):1–10.

Ahloy, J. and Hamman, J. R. (2019). Personality Traits and Endogenous Group Formation. *Source: Revue économique*, 70(6):999–1020.

Akerlof, G. A. and Kranton, R. E. (1997). Social Distance and Social Decisions. Technical Report 5.

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

Akerlof, G. A. and Kranton, R. E. (2002). Identity and Schooling: Some Lessons for the Economics of Education. *Journal of Economic Literature*, 40:1167–1201.

Alfonso-Costillo, A., Brañas-Garza, P., and López-Martín, M. C. (2022). Does the die-under-the-cup device exaggerate cheating? *Economics Letters*, 214.

Amir, A., Kogut, T., and Bereby-Meyer, Y. (2016). Careful cheating: People cheat groups rather than individuals. *Frontiers in Psychology*, 7(371):1–8.

Anvari, F., Wenzel, M., Woodyatt, L., and Haslam, S. A. (2019). The social psychology of whistleblowing: An integrated model. *Organizational Psychology Review*, 9(1):41–67.

Aquino, K. and Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6):1423–1440.

Aron, A., Aron, E., 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):596–612.

Artinger, F., Exadaktylos, F., Koppel, H., and Sääksvuori, L. (2010). Applying Quadratic Scoring Rule transparently in multiple choice settings: A note. *Working Paper*, (January):1–15.

Ashton, M. C. and Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2):150–166.

Ashton, M. C. and Lee, K. (2008). The prediction of Honesty-Humility-related criteria by the HEXACO and Five-Factor Models of personality. *Journal of Research in Personality*, 42(5):1216–1228.

Baader, M., Starmer, C., Tufano, F., and Gächter, S. (2024). Introducing IOS11 as an extended interactive version of the 'Inclusion of Other in the Self' scale to estimate relationship closeness. *Scientific Reports*, 14(1).

Baccini, E. and Hartmann, S. (2022). The Myside Bias in Argument Evaluation: A

Bayesian Model. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022*, pages 1512–1518.

Bäker, A. . and Mechtel, M. (2015). Peer Settings Induce Cheating on Task Performance.

Balliet, D., Wu, J., and De Dreu, C. K. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychological bulletin*, 140(6):1556–1581.

Barranti, M., Carlson, E. N., and Furr, R. M. (2016). Disagreement About Moral Character Is Linked to Interpersonal Costs. *Social Psychological and Personality Science*, 7(8):806–817.

Bartels, D. M. and Burnett, R. C. (2011). A group construal account of drop-in-the-bucket thinking in policy preference and moral judgment. *Journal of Experimental Social Psychology*, 47(1):50–57.

Beer, A. and Watson, D. (2008). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment*, 90(3):250–260.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.

Benistant, J., Galeotti, F., and Villeval, M. C. (2021). The Distinct Impact of Information and Incentives on Cheating The Distinct Impact of Information and Incentives on Cheating *. Technical report.

Benistant, J., Galeotti, F., and Villeval, M. C. (2022). Competition, information, and the erosion of morals. *Journal of Economic Behavior and Organization*, 204:148–163.

Beranek, B. and Castillo, G. (2022). Continuous Inclusion of Other in the Self. Technical report.

Berg, A. (2019). Identity in economics: a review.

Bernard, M., Hett, F., and Mechtel, M. (2016). Social identity and social free-riding. *European Economic Review*, 90:4–17.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, 132:59–72.

Bilancini, E., Boncinelli, L., Capraro, V., Celadin, T., and Di Paolo, R. (2020). "Do the right thing" for whom? An experiment on ingroup favouritism, group assorting and moral suasion. *Judgment and Decision Making*, 15(2):182–192.

Blanken, I., van de Ven, N., and Zeelenberg, M. (2015). A Meta-Analytic Review of Moral Licensing. *Personality and Social Psychology Bulletin*, 41(4):540–558.

Blokland, T. (2012). Blaming neither the undeserving poor nor the revanchist middle classes: A relational approach to marginalization. *Urban Geography*, 33(4):488–507.

Bone, J. E., McAuliffe, K., and Raihani, N. J. (2016). Exploring the motivations for punishment: Framing and country-level effects. *PLoS ONE*, 11(8).

Boone, C., Declerck, C., and Kiyonari, T. (2010). Inducing Cooperative Behavior among Proselfs versus Prosocials: The Moderating Role of Incentives and Trust. *Journal of Conflict Resolution*, 54(5):799–824.

Bose, N. and Sgroi, D. (2022). The role of personality beliefs and "small talk" in strategic behaviour. *PLoS ONE*, 17(9 September).

Brown-Iannuzzi, J. L., Lundberg, K. B., and McKee, S. E. (2021). Economic inequality and socioeconomic ranking inform attitudes toward redistribution. *Journal of Experimental Social Psychology*, 96.

Bussolo, M., Lebrand, M., and Torre, I. (2020). Feeling Poor, Feeling Rich, or Feeling Middle-Class An Empirical Investigation.

Cameron, L., Chaudhuri, A., Erkal, N., and Gangadharan, L. (2009). Propensities to engage in and punish corrupt behavior: Experimental evidence from Australia, India, Indonesia and Singapore. *Journal of Public Economics*, 93(7-8):843–851.

Castillo, G. (2021). Preference reversals with social distances. *Journal of Economic Psychology*, 86.

Castro Santa, J., Exadaktylos, F., and Soto-Faraco, S. (2018). Beliefs about others' intentions determine whether cooperation is the faster choice. *Scientific Reports*, 8(1):1–10.

Chae, J., Kim, K., Kim, Y., Lim, G., Kim, D., and Kim, H. (2022). Ingroup favoritism overrides fairness when resources are limited. *Scientific Reports*, 12(1).

Charness, G. (2000). Self-serving cheap talk: A test of aumann's conjecture. *Games and Economic Behavior*, 33(2):177–194.

Charroin, L., Fortin, B., Villeval, M. C., Boucher, V., Bramoullé, Y., Chen, Y., Cohn, A., Davidson, R., Fluet, C., Marchand, S., and Shearer, B. (2021). Homophily, Peer Effects, and Dishonesty Homophily, Peer Effects, and Dishonesty *. Technical report.

Chierchia, G. and Coricelli, G. (2015). The impact of perceived similarity on tacit coordination: Propensity for matching and aversion to decoupling choices. *Frontiers in Behavioral Neuroscience*, 9(JULY).

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., and Neuberg, S. L. (1997). Reinterpreting the Empathy-Altruism Relationship: When One Into One Equals Oneness. Technical report.

Conrads, J., Irlenbusch, B., Rilke, R. M., and Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, 34:1–7.

Cooper, J. (2019). Cognitive dissonance: Where we've been and where we're going. *International Review of Social Psychology*, 32(1).

Cooper, W. H. and Withey, M. J. (2009). The strong situation hypothesis. *Personality and Social Psychology Review*, 13(1):62–72.

Costa, P. (1992). Neo PI-R professional manual GWAS of Personality View project Lifespan and Intergenerational Effects of Childhood Malnutrition View project. Technical report.

Costa, P. T. and McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment: Volume 2 -*

Personality Measurement and Testing, pages 179–198. SAGE Publications Inc.

Coyne, I. and Bartram, D. (2002). Assessing the Effectiveness of Integrity Tests: A Review. *International Journal of Testing*, 2(1):15–34.

Crede, A.-K. and von Bieberstein, F. (2020). Reputation and lying aversion in the die roll paradigm: Reducing ambiguity fosters honest behavior. *Managerial and Decision Economics*, 41(4).

Crowe, M. L., Lynam, D. R., and Miller, J. D. (2018). Uncovering the structure of agreeableness from self-report measures. *Journal of Personality*, 86(5):771–787.

Currarini, S. and Mengel, F. (2016). Identity, homophily and in-group bias. *European Economic Review*, 90:40–55.

Dalton, R. (2016). Party identification and its implications. *Oxford Research Encyclopedia of Politics*.

de Dreu, C. K. (2010). Social value orientation moderates ingroup love but not outgroup hate in competitive intergroup conflict. *Group Processes & Intergroup Relations*, 13(6):701–713.

De Freitas, J., Thomas, K., DeScioli, P., and Pinker, S. (2019). Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences of the United States of America*, 116(28):13751–13758.

Deutchman, P., Bračič, M., Raihani, N., and McAuliffe, K. (2021). Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evolution and Human Behavior*, 42(1):12–20.

Devetag, G. and Ortmann, A. (2007). When and why? A critical survey on coordination failure in the laboratory. *Experimental Economics*, 10(3):331–344.

Dimant, E. (2019). Contagion of pro- and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73:66–88.

Drouvelis, M. and Georgantzis, N. (2019). Does revealing personality data affect prosocial behaviour? *Journal of Economic Behavior and Organization*, 159:409–420.

Du, H., Chen, A., Chi, P., and King, R. B. (2020). Preprint of "Income Inequality Reduces Civic Honesty".

Dungan, J. A., Young, L., and Waytz, A. (2019). The power of moral concerns in predicting whistleblowing decisions. *Journal of Experimental Social Psychology*, 85.

Easterbrook, M. J., Hadden, I. R., and Nieuwenhuis, M. (2019). Identities in context: How social class shapes inequalities in education. In *The Social Psychology of Inequality*, pages 103–121. Springer International Publishing.

Easterbrook, M. J., Kuppens, T., and Manstead, A. S. (2020). Socioeconomic status and the structure of the self-concept. *British Journal of Social Psychology*, 59(1):66–86.

Elbæk, C. T., Mitkidis, P., Aarøe, L., and Otterbring, T. (2023). Subjective socioeconomic status and income inequality are associated with self-reported morality across 67 countries. *Nature Communications*, 14(1).

Ellemers, N., Pagliaro, S., Barreto, M., and Leach, C. W. (2008). Is It Better to Be Moral

Than Smart? The Effects of Morality and Competence Norms on the Decision to Work at Group Status Improvement. *Journal of Personality and Social Psychology*, 95(6):1397–1410.

Fagbenro, D. A. (2019). Personality Traits and Attitude toward Corruption among Government Workers. *Psychology and Behavioral Science International Journal*, 11(1).

Falk, A. and Zimmermann, F. (2024). Attention and Dread: Experimental Evidence on Preferences for Information. *Management Science*, 70(10):7090–7100.

Fan, C. S., Wei, X., Wu, J., and Zhang, J. (2022). Observability and peer effects: Theory and evidence from a field experiment. *Journal of Economic Behavior and Organization*, 200:847–867.

Fehr, D., Kübler, D., and Danz, D. (2008). Information and Beliefs in a Repeated Normal-form game. *Philosophy of Information*, (3627):551–577.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise-an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Fries, T., Gneezy, U., Kajackaite, A., and Parra, D. (2021). Observability and lying. *Journal of Economic Behavior and Organization*, 189:132–149.

Gächter, S. and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595):496–499.

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).

Galeotti, F., Rilke, R. M., and Verrina, E. (2024). Beliefs and Group Dishonesty: The Role of Strategic Interaction and Responsibility. Technical report.

Gibson, R., Tanner, C., and Wagner Alexander F. (2013). Preferences for Truthfulness: }Heterogeneity Among and Within Individuals. *American Economic Review*, 103(1):532–548.

Gino, F., Ayal, S., and Ariely, D. (2009). Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel. Technical Report 3.

Gino, F., Ayal, S., and Ariely, D. (2012). Self-Serving Altruism? When Unethical Actions That Benefit Others Do Not Trigger Guilt. Technical report.

Giorgetta, C., Grecucci, A., Graffeo, M., Bonini, N., Ferrario, R., and Sanfey, A. G. (2021). Expect the Worst ! Expectations and Social Interactive Decision Making. *Brain Sciences*, 11(572).

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Goldstein, N. J. and Cialdini, R. B. (2007). The spyglass self: A model of vicarious self-perception. *Journal of Personality and Social Psychology*, 92(3):402–417.

Greenberg, S. and Org, C. (2018). Calibration Scoring Rules for Practical Prediction Training. Technical report.

Gries, T., Müller, V., and Jost, J. T. (2022). The Market for Belief Systems: A Formal

Model of Ideological Choice. *Psychological Inquiry*, 33(2):65–83.

Grigoryan, L. (2020). Crossed categorization outside the lab: Findings from a factorial survey experiment. *European Journal of Social Psychology*, 50(5):983–1000.

Grigoryan, L., Seo, S., Simunovic, D., and Hofmann, W. (2023). Helping the ingroup versus harming the outgroup: Evidence from morality-based groups. *Journal of Experimental Social Psychology*, 105.

Gross, J. and De Dreu, C. K. (2021). Rule Following Mitigates Collaborative Cheating and Facilitates the Spreading of Honesty Within Groups. *Personality and Social Psychology Bulletin*, 47(3):395–409.

Gross, J., Leib, M., Offerman, T., and Shalvi, S. (2018). Ethical Free Riding: When Honest People Find Dishonest Partners. *Psychological Science*, 29(12):1956–1968.

Gueguen, N., Jacob, C., and Martin, A. (2009). Mimicry in social interaction: Its effect on human judgment and behavior. *European Journal of Social Sciences*, 8(2):253–259.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M., Lagos, P., Norris, E., Ponarin, and B. Puranen et al. (eds.) (2020). World Values Survey: Round Seven – Country-Pooled Datafile. Technical report, JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Hauge, L. (2007). Identity and Place: A Critical Comparison of Three Identity Theories.

Hauser, O. P., Kraft-Todd, G. T., Rand, D. G., Nowak, M. A., and Norton, M. I. (2021). Invisible inequality leads to punishing the poor and rewarding the rich. *Behavioural Public Policy*, 5(3):333–353.

Hermann, D. and Ostermaier, A. (2018). Be close to me and I will be honest How social distance influences honesty.

Hershcovis, M. S., Neville, L., Reich, T. C., Christie, A. M., Cortina, L. M., and Shan, J. V. (2017). Witnessing wrongdoing: The effects of observer power on incivility intervention in the workplace. *Organizational Behavior and Human Decision Processes*, 142:45–57.

Hewston, M., Rubin, M., and Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, (53):575–604.

Heyman, T., Vankrunkelsven, H., Voorspoels, W., White, A., Storms, G., and Verheyen, S. (2020). When Cheating is an Honest Mistake: A Critical Evaluation of the Matrix Task as a Measure of Dishonesty. *Collabra: Psychology*, 6(1).

Hilbig, B. E., Hessler, C. M., Thielmann, I., Wührl, J., and Zettler, I. (2015). What lies beneath: How the distance between truth and lie drives dishonesty. *Personality and Individual Differences*, 80(2):263–266.

Hilbig, B. E., Zettler, I., Leist, F., and Heydasch, T. (2013). It takes two: Honesty-Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, 54(5):598–603.

Hoffmann, T. (2013). The Effect of Belief Elicitation on Game Play. pages 1–26.

Holt, C. A. and Laury, S. K. (2002). Risk Aversion and Incentive Effects. Technical

Report 5.

Hughes, B. T., Flournoy, J. C., and Srivastava, S. (2020). Is Perceived Similarity More Than Assumed Similarity?: An Interpersonal Path to Seeing Similarity Between Self and Others. *Journal of Personality and Social Psychology*, 121(1):184–200.

Hyndman, K., Terracol, A., and Vaksmann, J. (2013). Beliefs and (In)Stability in Normal-Form Games. (47221).

Inglehart, R. (2000). Culture and Democracy. In *Culture Matters: How Values Shape Human Progress*, pages 80–97. New York: Basic Books.

INSEE Références (2021). La France et ses territoires. Technical report.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2025). The Origins and Consequences of Affective Polarization in the United States. 53:29.

Jansson, F. (2015). What games support the evolution of an ingroup bias? *Journal of Theoretical Biology*, 373:100–110.

Jansson, F. and Eriksson, K. (2015). Cooperation and shared beliefs about trust in the assurance game. *PLoS ONE*, 10(12):1–13.

John, O., Naumann, L., and Soto, C. (2008). *Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues.* Guilford Press, 3rd edition.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90(12).

Jordan, P. J., Troth, A. C., and Yan, H. (2024). Objective and subjective measurement in applied business settings: Improving research in organizations. *Australian Journal of Management*.

Kajonius, P. J. and Dåderman, A. M. (2014). Exploring the relationship between honesty-humility, the big five, and liberal values in Swedish students. *Europe's Journal of Psychology*, 10(1):104–117.

Kaluza, B., Institute, J. S., Kaminka, G., Tambe, M., Kaluža, B., and Kaminka, G. A. (2012). Detection of suspicious behavior from a sparse set of multiagent interactions. Technical report.

Kang, P., Burke, C. J., Tobler, P. N., and Hein, G. (2021). Why we learn less from observing outgroups. *Journal of Neuroscience*, 41(1):144–152.

Kaushik, M., Singh, V., and Chakravarty, S. (2021). Rewards, Detection and Dishonesty: Experimental Evidence from India. *SSRN Electronic Journal*.

Kim, J. E. and Tsvetkova, M. (2021). Cheating in online gaming spreads through observation and victimization. *Network Science*, 9(4):425–442.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms Make Preferences Social. *Journal of the European Economic Association*, 14(3):608–638.

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2019). Intuitive Honesty Versus Dishonesty: Meta-Analytic Evidence. *Perspectives on Psychological Science*, 14(5):778–796.

Kocher, M., Martinsson, P., and Visser, M. (2012). Social background, cooperative behavior, and norm enforcement. *Journal of Economic Behavior and Organization*, 81(2):341–354.

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, 64(9):3995–4008.

Korbel, V. (2016). Do we lie in groups? An experimental evidence. *Applied Economic Letters*, 24(15):1107–1111.

Kraus, M. W., Piff, P. K., and Keltner, D. (2011). Social class as culture: The convergence of resources and rank in the social realm. *Current Directions in Psychological Science*, 20(4):246–250.

Kreps, D. M. (1992). *Game Theory and Economic Modelling*. Oxford.

Kroher, M. and Wolbring, T. (2015). Social control, social learning, and cheating: Evidence from lab and online experiments on dishonesty. *Social Science Research*, 53:311–324.

Ladley, D., Wilkinson, I., and Young, L. (2015). The impact of individual versus group rewards on work group performance and cooperation: A computational social science approach. *Journal of Business Research*, 68(11):2412–2425.

Lane, T. (2023). The strategic use of social identity CeDEx Discussion Paper Series. Technical report.

Larrouy, L. and Lecouteux, G. (2017). Mindreading and endogenous beliefs in games. *Journal of Economic Methodology*, 24(3):318–343.

Le Coq, C., Tremewan, J., and Wagner, A. K. (2015). On the effects of group identity in strategic environments. *European Economic Review*, 76:239–252.

Lee, J. J., Hardin, A. E., Parmar, B., and Gino, F. (2019). The interpersonal costs of dishonesty: How dishonest behavior reduces individuals' ability to read others' emotions. *Journal of Experimental Psychology: General*, 148(9):1557–1574.

Leib, M., Köbis, N., Soraperra, I., Weisel, O., and Shalvi, S. (2021). Collaborative Dishonesty: A Meta-Analytic Review. *Psychological Bulletin*, 147(12):1241–1268.

Leibbrandt, A., López-Pérez, R., and Spiegelman, E. (2023). Reciprocal, but inequality averse as well? Mixed motives for punishment and reward. *Journal of Economic Behavior and Organization*, 210:91–116.

Leidner, B., Castano, E., Zaiser, E., and Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin*, 36(8):1115–1129.

Lönnqvist, J. E., Ilmarinen, V. J., and Verkasalo, M. (2021). Who likes whom? The interaction between perceiver personality and target look. *Journal of Research in Personality*, 90.

Loustau, T., Glassman, J., Martin, J. W., Young, L., and McAuliffe, K. (2024). The impact of group membership on punishment versus partner rejection. *Scientific Reports*, 14(1).

Lutz, G. and Lauener, L. (2020). Measuring party affiliation. Technical report, Lausanne: Swiss Centre of Expertise in the Social Sciences (FORS)., Lausanne.

Macků, K., Caha, J., Pászto, V., and Tuček, P. (2020). Subjective or objective? How objective measures relate to subjective life satisfaction in Europe. *ISPRS International Journal of Geo-Information*, 9(5).

Magni, G. (2021). Economic inequality, immigrants and selective solidarity: From perceived lack of opportunity to in-group favoritism.

Mann, H., Garcia-Rada, X., Houser, D., and Ariely, D. (2014). Everybody else is doing it: Exploring social transmission of lying behavior. *PLoS ONE*, 9(10).

Manstead, A. S. (2018). The psychology of social class: How socioeconomic status impacts thought, feelings, and behaviour. *British Journal of Social Psychology*, 57(2):267–291.

Martin, R. A. (2015). *Perceived and Actual Similarity as Predictors of Self-Disclosure and Perceived Understanding at Zero Acquaintance*. PhD thesis.

Martinangeli, A. F. and Martinsson, P. (2020). We, the rich: Inequality, identity and cooperation. *Journal of Economic Behavior and Organization*, 178:249–266.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people. *Journal of Marketing Research*, 45(6):633–644.

McFerran, B., Aquino, K., and Duffy, M. (2010). How Personality and Moral Identity Relate to Individuals' Ethical Ideology. *Business Ethics Quarterly*, 20(1):35–56.

Mendoza, S. A., Lane, S. P., and Amodio, D. M. (2014). For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game. *Social Psychological and Personality Science*, 5(6):662–670.

Meyners, J., Barrot, C., Becker, J. U., and Goldenberg, J. (2017). The role of mere closeness: How Geographic proximity affects social influence. *Journal of Marketing*, 81(5):49–66.

Michaeli, M. (2020). Grouping, in-group bias and the cost of cheating. *Games and Economic Behavior*, 121:90–107.

Molho, C., De Petrillo, F., Garfield, Z. H., and Slewe, S. (2024). Cross-societal variation in norm enforcement systems.

Moss, R. H., Kelly, B., Bird, P. K., and Pickett, K. E. (2023). Examining individual social status using the MacArthur Scale of Subjective Social Status: Findings from the Born in Bradford study. *SSM - Population Health*, 23.

OECD (2024). OECD Survey on Drivers of Trust in Public Institutions – 2024 Results: Building Trust in a Complex Policy
}Environment. Technical report, OECD Publishing, Paris.

Offerman, T., Sonnemans, J., van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies*, 76(4):1461–1489.

Owuamalam, C. K., Rubin, M., Spears, R., and Weerabangsa, M. M. a. (2017). Why Do

People from Low-Status Groups Support Class Systems that Disadvantage Them? A Test of Two Mainstream Explanations in Malaysia and Australia. *Journal of Social Issues*, 73(1):80–98.

Panagopoulos, C., Leighley, J. E., and Hamel, B. T. (2017). Are Voters Mobilized by a 'Friend-and-Neighbor'on the Ballot? Evidence from a Field Experiment. *Political Behavior*, 39(4):865–882.

Pansini, R., Campennì, M., and Shi, L. (2018). Asymmetric use of punishment in socioeconomic segregated societies leads to an unequal distribution of wealth. Technical report.

Proto, E., Rustichini, A., Deyoung, C., Friebel, G., Grimalda, G., Isoni, A., Loomes, G., Manzini, P., Mariotti, M., Miller, J., Oswald, A., and Stewart, N. (2014). Cooperation and Personality. Technical report.

Pulfrey, C., Durussel, K., and Butera, F. (2018). The good cheat: Benevolence and the justification of collective cheating. *Journal of Educational Psychology*, 110(6):764–784.

Rantakari, H. (2023). How to reward honesty? *Journal of Economic Behavior and Organization*, 207:129–145.

Régner, I. and Monteil, J.-M. (2007). Low-and high-socioeconomic status students preference for ingroup comparisons and their underpinning ability expectations. *Revue Internationale de Psychologie Sociale*, 20(1):87–104.

Renger, D., Lohmann, J. F., Renger, S., and Martiny, S. E. (2024). Socioeconomic status and self-regard income predicts self-respect over time. *Social Psychology*, 55(1):12–24.

Rijnks, R. H. and Strijker, D. (2013). Spatial effects on the image and identity of a rural area. *Journal of Environmental Psychology*, 36:103–111.

Robalo, P., Schram, A., and Sonnemans, J. (2017). Other-regarding preferences, in-group bias and political participation: An experiment. *Journal of Economic Psychology*, 62:130–154.

Rothstein, B. (2011). Anti-corruption: The indirect 'big bang' approach. *Review of International Political Economy*, 18(2):228–250.

Rothstein, B. and Eek, D. (2009). Political Corruption and Social Trust. *Rationality and Society*, 21(1):81–112.

Rubin, M., Badea, C., and Jetten, J. (2014). Low status groups show in-group favoritism to compensate for their low status and compete for higher status. *Group Processes & Intergroup Relations*, 17(5):563–576.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under "almost common knowledge". *American Economic Review*, 79(3):385–391.

Rubinstein, A. and Salant, Y. (2016). Isn't everyone like me?": On the presence of self-similarity in strategic interactions. *Judgment and Decision Making*, 11(2):168–173.

Ruch, W., Bruntsch, R., and Wagner, L. (2017). The role of character traits in economic

games. *Personality and Individual Differences*, 108:186–190.

Rullo, M., Monaco, S., Giannini, F., Livi, S., and Presaghi, F. (2019). In the name of truth: People's reactions to ingroup and outgroup members who self-disclose a severe error. *Social Science Journal*, 56(3):421–424.

Rullo, M., Presaghi, F., Baldner, C., Livi, S., and Butera, F. (2024). Omertà in intragroup cheating: The role of ingroup identity in dishonesty and whistleblowing. *Group Processes and Intergroup Relations*, 27(1):41–61.

Rustichini, A. (2009). Neuroeconomics: what have we found, and what should we search for.

Rustichini, A., DeYoung, C. G., Anderson, J. E., and Burks, S. V. (2016). Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation. *Journal of Behavioral and Experimental Economics*, 64:122–137.

Ruzzier, C. A. and Woo, M. D. (2023). Discrimination with inaccurate beliefs and confirmation bias. *Journal of Economic Behavior and Organization*, 210:379–390.

Ryvkin, D., Serra, D., and Tremewan, J. (2017). I paid a bribe: An experiment on information sharing and extortionary corruption. *European Economic Review*, 94:1–22.

Schiller, B., Baumgartner, T., and Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior*, 35(3):169–175.

Schram, A., Zheng, J. D., and Zhuravleva, T. (2022). Corruption: A cross-country comparison of contagion and conformism. *Journal of Economic Behavior and Organization*, 193:497–518.

Sgroi, D., Yeo, J., and Zhuo, S. (2021). Ingroup Bias with Multiple Identities: The Case of Religion and Attitudes Towards Government Size. Technical report.

Shalvi, S., Dana, J., Handgraaf, M. J., and De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190.

Siniver, E., Tobol, Y., and Yaniv, G. (2022). Collective Punishment and Cheating in the Die-Under-the-Cup Task. *Experimental Psychology*, 69(1):40–45.

Skyrms, B. (2003). *The Stag Hunt and the Evolution of Social Structure*. Number 1. Cambridge University Press, Cambridge.

Sosa, M. and Maoret, M. (2023). Close to Me: The Impact of the Interplay of Physical and Social Proximity on Dyadic Collaboration Effectiveness. Technical report.

Stahl, D. and Huyck, J. V. (2002). Learning conditional behavior in similar stag hunt games. (January).

Steinel, W., Valtcheva, K., Gross, J., Celse, J., Max, S., and Shalvi, S. (2022). (Dis)honesty in the face of uncertain gains or losses. *Journal of Economic Psychology*, 90.

Tajfel, H. and Turner, J. C. (1979). An integrative theory of intergroup conflict. In *The social psychology of intergroup relations*, pages 33–37. Brooks/Cole, Monterey, CA.

Thielmann, I., Akrami, N., Babarović, T., Belloch, A., Bergh, R., Chirumbolo, A., Čolović, P., de Vries, R. E., Dostál, D., Egorova, M., Gnisci, A., Heydasch, T., Hilbig, B. E., Hsu, K. Y., Izdebski, P., Leone, L., Marcus, B., Međedović, J., Nagy, J., Parshikova, O., Perugini, M., Petrović, B., Romero, E., Sergi, I., Shin, K. H., Smederevac, S., Šverko, I., Szarota, P., Szirmák, Z., Tatar, A., Wakabayashi, A., Wasti, S. A., Záškodná, T., Zettler, I., Ashton, M. C., and Lee, K. (2020). The HEXACO–100 Across 16 Languages: A Large-Scale Test of Measurement Invariance. *Journal of Personality Assessment*, 102(5):714–726.

Thielmann, I., Hilbig, B. E., Klein, S. A., Seidl, A., and Heck, D. W. (2024). Cheating to benefit others? On the relation between Honesty-Humility and prosocial lies. *Journal of Personality*, 92(3):870–882.

Thomas, G. O., Poortinga, W., and Sautkina, E. (2016). The Welsh Single-Use Carrier Bag Charge and behavioural spillover. *Journal of Environmental Psychology*, 47(2880):126–135.

Thomas, K. A., DeScioli, P., Haque, O. S., and Pinker, S. (2014). The Psychology of Coordination and Common Knowledge. *Journal of Personality and Social Psychology*, 107(4):657–676.

Tobol, Y., Siniver, E., and Yaniv, G. (2020). Do tightwads cheat more? Evidence from three field experiments. *Journal of Economic Behavior and Organization*, 180:148–158.

Tsvetkova, M. and Macy, M. W. (2015). The social contagion of antisocial behavior. *Sociological Science*, 2:36–49.

Van Assche, J., Politi, E., Van Dessel, P., and Phalet, K. (2020). To punish or to assist? Divergent reactions to ingroup and outgroup members disobeying social distancing. *British Journal of Social Psychology*, 59(3):594–606.

van de Ven, J. and Villeval, M. C. (2015). Dishonesty under scrutiny. *Journal of the Economic Science Association*, 1(1):86–99.

Van De Walle, S. (2008). *Perceptions of corruption as distrust? Cause and effect in attitudes toward government*. Number June.

Van Huyck, J., Viriyavipart, A., and Brown, A. L. (2018). When less information is good enough: experiments with global stag hunt games. *Experimental Economics*, 21(3):527–548.

Van Huyck, J. B., Battalio, R. C., and Beil, R. O. (1990). Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure. *American Economic Review*, 80(1):234–248.

van Oosten, S. (2025). The Importance of In-group Favoritism in Explaining Voting for PRRPs: A Study of Minority and Majority Groups in France, Germany and the Netherlands. Technical report, European Center for Populism Studies, Brussels.

Volk, S., Thöni, C., and Ruigrok, W. (2011). Personality, personal values and cooperation preferences in public goods games: A longitudinal study. *Personality and Individual Differences*, 50(6):810–815.

Waytz, A., Dungan, J., and Young, L. (2013). The whistleblower's dilemma and the fairness-loyalty tradeoff. *Journal of Experimental Social Psychology*, 49(6):1027–1033.

Weiner, D. S. and Laurent, S. M. (2021). The (Income-Adjusted) Price of Good Behavior: Documenting the Counter-Intuitive, Wealth-Based Moral Judgment Gap. *Journal of Experimental Psychology: General*, 150(3):484–506.

Weisel, O. and Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences of the United States of America*, 112(34):10651–10656.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, 115(2):157–168.

Windrich, I., Kierspel, S., Neumann, T., Berger, R., and Vogt, B. (2024). Enforcement of Fairness Norms by Punishment: A Comparison of Gains and Losses. *Behavioral Sciences*, 14(1).

Winter, F. and Zhang, N. (2018). Social norm enforcement in ethnically diverse communities. 115(11):2722–2727.

Wu, J., Balliet, D., and Van Lange, P. A. (2016). Reputation, Gossip, and Human Cooperation.

Zhao, K. and Smillie, L. D. (2015). The Role of Interpersonal Traits in Social Decision Making: Exploring Sources of Behavioral Heterogeneity in Economic Games. *Personality and Social Psychology Review*, 19(3):277–302.

Zhou, L., Su, C., Sun, X., Zhao, X., and Choo, K. K. R. (2018). Stag hunt and trust emergence in social networks. *Future Generation Computer Systems*, 88:168–172.

Chapter 3. Socioeconomic Distance and the Selectivity in Punishment and Reward

Irving Argaez Corona[a], Béatrice Boulu-Reshef[b], Jean-Christophe Vergnaud[a,c]

[a]Centre d'Économie de la Sorbonne (CES), Université Paris 1 Panthéon-Sorbonne, France
[b]CY Cergy Paris Université (THEMA), France
[c]Centre National de la Recherche Scientifique (CNRS), France

Abstract

Do we punish distant others more harshly and reward close others more readily? We study these decisions in an iterated online Die-under-the-cup (DUTC) task with $n = 720$ participants randomly assigned to fixed roles: Die-rollers (P1) and Observers (P2). Over 24 rounds, each P2 is paired with two 12-round blocks of P1s-one socially close and one socially distant—under one of three treatments: Punishment, Reward, or Mixed. P2 privately observes both P1's true and reported values, and can apply -€2 penalty or a +€2 reward that does not affect P2's own payoff. Social closeness is modelled through socioeconomic status (SES) using (1) an objective measure based on average income in the participant's locality of residence and (2) a subjective measure based on self-reported income relative to the locality's average. We construct an observer-level proxy for *suspicious reporting* using excess self-serving reports relative to fair odds within the 12 reports seen by P2 per SES and test whether social closeness moderates enforcement. Findings show that punishment increases with the suspicion proxy but is weaker for close counterparts, while rewards are granted more often for non-suspicious reports. Third-party penalties and rewards are selective across SES: at comparable suspicion levels, low-SES counterparts are punished more than medium/high-SES counterparts, yet they also receive more rewards for non-suspicious reports. Moreover, we observe that the objective measure of closeness is linked to overall leniency toward close counterparts, whereas the subjective measure shifts reactions to the suspicion proxy: weaker for close counterparts and stronger for distant ones. Overall, our study reveals that counterpart SES shapes responses to reported outcomes.

This project benefitted from funding from the *Centre d'Économie de la Sorbonne (CES)*'s 2025 campaign. The project received approval from the Institutional Review Board of Paris School of Economics (decision 2025-030). The pre-registered protocol can be found in this link, as well as in the Appendix, for peer-review purposes. We warmly thank **Maxim Frolov**, computer engineer at *Laboratoire d'Economie Expérimentale de Paris (LEEP)*, for his support in programming the experiment and organising the sessions.

Résumé

Punissons-nous davantage les personnes socialement éloignées et récompensons-nous plus facilement les proches ? Nous étudions ces décisions dans une tâche en ligne itérative de type *Die-under-the-cup (DUTC)* avec $n = 720$ participants, assignés aléatoirement à des rôles fixes : Lanceurs de dé (P1) et Observateurs (P2). Sur 24 manches, chaque P2 est apparié à deux blocs de 12 P1—l'un socialement proche, l'autre socialement distant—selon l'un de trois traitements : Punition, Récompense ou Mixte. P2 observe en privé la valeur réelle et la valeur déclarée par P1, et peut appliquer une pénalité de -2 ou une récompense de $+2$ qui n'affecte pas son propre gain. La proximité sociale est modélisée via le statut socio-économique (SES) en utilisant (1) une mesure objective fondée sur le revenu moyen de la localité de résidence du participant et (2) une mesure subjective basée sur le revenu auto-déclaré relativement à la moyenne de la localité. Nous construisons, au niveau de l'observateur, un proxy de déclaration suspecte à partir de l'excès de déclarations auto-avantageuses par rapport aux chances équitables parmi les 12 déclarations vues par P2 pour chaque SES, et testons si la proximité sociale modère l'application des sanctions. Les résultats montrent que la punition augmente avec le proxy de suspicion mais est plus faible envers les proches, tandis que les récompenses sont plus souvent accordées pour des déclarations non suspectes. Les pénalités et récompenses de tiers sont sélectives selon le SES : à niveaux de suspicion comparables, les P1 de bas SES sont davantage punis que ceux de SES moyen/élevé, mais reçoivent aussi plus de récompenses pour des déclarations non suspectes. En outre, la mesure objective de proximité est associée à une indulgence générale envers les proches, tandis que la mesure subjective module la réaction au proxy de suspicion : plus faible envers les proches et plus forte envers les distants. Globalement, notre étude révèle que le SES du partenaire façonne les réponses aux déclarations observées.

3.1 Introduction

Individual decisions to cheat often transcend personal consequences and impact broader social outcomes. Evidence suggests that seeing similar others cheat can increase dishonesty among observers (Bicchieri et al., 2022; Dimant, 2019; Jordan et al., 2024; Kang et al., 2021). In this sense, this paper addresses two questions and tests them experimentally: (1) when observers evaluate reported outcomes that differ in how suspicious they look, do they punish suspicious reports and reward non-suspicious reports?, and (2) does social closeness shape these reactions? We measure suspicion with an Observer-level proxy for cheating, vary social closeness within participants using objective and subjective measures of socioeconomic status (SES), and estimate how punishment and reward respond to the suspicion proxy when paired with close versus distant pairs.

Prior research shows that punishment choices depend on perceived harm and inequality and are shaped by in-group preferences (Bone et al., 2016; Deutchman et al., 2021; Rullo et al., 2024). Recent papers document that support for sanctions is heterogeneous, with in-group loyalty reducing willingness to sanction even under clear norm violations (Loustau et al., 2024; van Oosten, 2025), especially when sanctions threaten group status (Hershcovis et al., 2017; Leib et al., 2021; Rullo et al., 2019). At the same time, self-image concerns and the wish to appear trustworthy deter dishonesty (Barranti et al., 2016; Blanken et al., 2015; Fischbacher and Föllmi-Heusi, 2013; Mazar et al., 2008), with internalised norms operating most strongly within in-groups (Grigoryan et al., 2023; Kimbrough and Vostroknutov, 2016). These findings motivate testing whether social closeness alters how Observers react to suspicious versus non-suspicious reports.

The study examines these questions in a repeated online Die-under-the-Cup (DUTC) task where Observers (P2) rotate with 24 counterparts in two 12-round blocks, one with socially close and one with socially distant counterparts. In the setting, Observers see the true and the reported values from Die-rollers (P1), with true values being anonymous to the experimenters. Participants are randomly assigned to one of three treatments: *Punish* (P2 may impose a -2€ penalty to P1), *Reward* (P2 may provide a +2€ reward to P1), or *Mixed* (P2 is randomly assigned to one of three possible options: observe, punish, or reward). The design allows to capture responses to both suspicious and non-suspicious reports, as well as comparisons across SES tiers (i.e., low, medium, high).

We construct an Observer-level proxy for suspicious reporting by comparing, within each 12-round block, the frequency of each reported face to the uniform benchmark. We then weight any excess by whether the face is payoff-maximising or payoff-reducing, assigning a round-level suspicion signal. Moreover, we make Observers' payoffs independent to their decisions to punish or reward, in order to isolate loss-recovery or reciprocity motives in their decisions.

Social closeness is defined via socioeconomic status (SES) using an objective measure (average income in the participant's department of residence) and a subjective measure (self-reported income relative to the departmental average). This logic follows existing evidence on how inequality and SES similarity shape interpersonal judgement and behaviour (Elbæk et al., 2023; Martinangeli and Martinsson, 2020; Moss et al., 2023)).

The papers' contributions are threefold. First, we introduce a transparent, observer-level

suspicion proxy that preserves the DUTC's inference logic while allowing clean tests of how Observers react to suspicious versus non-suspicious reports. Second, we analyse third-party sanctions and rewards jointly, rather than focusing on punishment alone. Third, we identify how socioeconomic closeness moderates these reactions using both objective and subjective measures of SES.

Our results show that decisions to punish and reward are robustly associated with both the suspicion proxy and the counterpart's socioeconomic context. Punishment rises with suspicion but is attenuated for objectively close counterparts, consistent with in-group leniency. Moreover, punishment is also socioeconomically selective as low-SES counterparts are punished more than medium or high-SES ones for the same level of suspicion. Conversely, Observers reward non-suspicious reports more often, with low-SES counterparts receiving more rewards and high-SES ones receiving the least. We also find that subjective closeness does not shift average punishment or reward levels, but instead changes sensitivity to suspicion, with weaker reactions for close counterparts and stronger for distant ones. Overall, our study shows decisions to punish and reward are selective, shaped jointly by suspicions of cheating and by social closeness to the counterparts.

The paper proceeds as follows: Section 3.2 elaborates on the existing literature, the theoretical framework for the paper and its contributions. Section 3.3 presents the experimental design and procedures. Section 3.4 reports the descriptive statistics and Section 3.5 reports the results. Finally, Section 3.6 concludes.

3.2 Theoretical background and related literature

This study takes from and contributes to three key areas of the literature. In this section, we outline the theoretical foundations and empirical findings that inform our approach, where our study sits and how this theoretical background helps in testing our hypotheses.

3.2.1 Inferring cheating in the DUTC paradigm

A central challenge in dishonesty research is that, by design, many paradigms do not adjudicate cheating at the individual level. This is the case with the Die-under-the-cup (DUTC) task, where participants privately observe a die outcome and self-report for payoff; cheating is therefore inferred statistically from how reported outcomes deviate from the uniform distribution. This logic is used throughout the DUTC literature: truthful behaviour produces an approximately uniform distribution of reports and systematic excess of payoff-maximising reports signals dishonesty in aggregate terms (Fischbacher and Föllmi-Heusi, 2013; Kocher et al., 2018).

Inferential estimation is common to lab and online applications of the DUTC, where cheating is almost always inferred from report distributions (Hermann and Ostermaier, 2018; Kroher and Wolbring, 2015; Siniver et al., 2022). Moreover, evidence proves that this feature upholds even across modified versions of the paradigm, either by changing feedback (Kroher and Wolbring, 2015), inducing competition settings (Benistant et al., 2022) or introducing sanction schemes (Siniver et al., 2022). These adaptations can inflate or reduce over–reporting, whilst preserving the core logic that cheating is inferred from distribution and not from personalised behaviour.

While this approximation has made the DUTC a leading paradigm to experimentally detect dishonesty, it constrains how cheating is interpreted. A complementary literature contrasts aggregate with individual detection, showing that DUTC paradigms identify cheating only at the group level, whereas individual-verification and anomaly-scoring approaches can flag specific cheaters, thereby corroborating, rather than replacing, DUTC-based inference (Heyman et al., 2020; Kaluza et al., 2012). These studies suggest that aggregate and individual approaches illustrate different facets of dishonesty and that careful, context-sensitive inference remains necessary.

Our design follows this logic. True rolls are shown to P2 and subsequently deleted from the master data, hence not observed by experimenters. We therefore construct an Observer-level, signal-based proxy for suspicious reporting by flagging excess frequency on self-serving faces relative to fair odds within each 12-report block. We then study how penalties and rewards respond to this proxy and whether socioeconomic closeness moderates these responses. This preserves the DUTC's distributional inference while enabling clean tests of Observer behaviour across socioeconomic contexts.

3.2.2 Behavioural responses: punishment and reward

Evidence shows that group boundaries shape punitive and reward choices. Observers punish out-group violators more and are more lenient toward in-group members (Loustau et al., 2024; Van Assche et al., 2020); in-group favouritism and out-group discrimination both enter norm responses (Schiller et al., 2014). Enforcement also reflects concerns about harm and inequality (Leibbrandt et al., 2023) and is stronger when group outcomes are salient (Windrich et al., 2024). Across settings, social conditions, power dynamics, and background heterogeneity shift enforcement thresholds (Hershcovis et al., 2017; Kocher et al., 2012; Molho et al., 2024; Winter and Zhang, 2018). The common result is asymmetry: the same transgression draws harsher action against out-groups than in-groups (Rullo et al., 2019,2; Van Assche et al., 2020; Winter and Zhang, 2018).

However, punishment is only one lever, as rewards can complement or even substitute sanctions. In public-goods environments, punishment opportunities raise contributions, and well-designed rewards can stabilise cooperation; which lever dominates depends on the strategic context and prevailing norms (Kocher et al., 2012). Outside the lab, enforcement presence and reward salience curb dishonesty (Kaushik et al., 2021; Rantakari, 2023). Group-level reward schemes can outperform individual or mixed schemes by promoting cooperation (Ladley et al., 2015).

Our study builds directly on these insights and tackles enforcement across social distance as a central challenge in real-life interactions. First, Observers in our design see the Die-roller's report and decide whether to punish (-2€) or reward (+2€) with no impact to their own payoffs. By removing reciprocity, the design not only isolates the observed behaviour, but also eliminates personal consequences on decisions. Second, we focus on social closeness by varying socioeconomic (dis)similarity between participants. This allows us to propose a model that combines punishment and reward in a single enforcement framework across socioeconomic differences, all while preserving a clean research environment that allows for anonymity from

the researchers.

3.2.3 Socioeconomic status and selective attitudes

A large body of work shows that socioeconomic status (SES) is not merely a background descriptor but a significant social mechanism: it shapes self–concept, judgment of others and the way individuals relate to others. Socialisation between individuals with different SES produces distinct behaviours and social appraisals (Manstead, 2018), as people attach substantial psychological importance to SES–linked identities (Easterbrook et al., 2020). SES also organises social perceptions and interactions in different ways: Kraus et al. (2011) show that cues on social class systematically alter cognition and interpersonal behaviour; Chae et al. (2022) claim that resource scarcity fosters in–group favouritism, often at the expense of fairness toward outsiders; while Renger et al. (2024) find SES divergence transforms how norm violations are perceived and confronted, mainly related to unequal treatment towards lower-income individuals.

In this sense, Hauser et al. (2021) shows that when income distribution is unknown, observers tend to reward the rich for large, absolute contributions and punish the poor for small ones; when inequality is revealed, judgments flip to relative standards-punishing the rich for low percentage contributions and rewarding the poor for high percentage ones. More generally, identical transgressions targeting wealthy versus poor agents are judged through different lenses: poorer targets are often seen as less immoral for the same violation and as more prosocial for the same good deed (Weiner and Laurent, 2021). Similarly, evidence on civic honesty is mixed but equally status–driven: a lost–wallet experiment correlates higher income with lower civic honesty (Du et al., 2020), a large cross–national survey suggests that lower SES can go along with stronger ethical concerns, challenging simple "low–SES, low ethicality" stereotypes (Elbæk et al., 2023), and SES ranking and perceived inequality causally shape redistribution attitudes through appraisals of fairness (Brown-Iannuzzi et al., 2021).

Furthermore, selective enforcement depends on who holds the power to sanction. For instance, Pansini et al. (2018) analyse class–segregated settings where the rich can punish the poor, finding that the rich cooperate less, punish more and end up earning roughly twice as much as the poor. These patterns echo broader preferences skewed against lower-income individuals, while recognising how everyday interactions and institutions reproduce exclusion (Blokland, 2012). They also explain why many people prefer to self–identify as "middle class" even when objective indicators place them low in the economic ladder (Bussolo et al., 2020).

This literature yields clear and testable implications for our setting. First, if distance reduces empathy and heightens discriminatory patterns, socioeconomic distance should amplify punishment when an observer perceives suspicious and payoff–maximising reports. Second, socioeconomic closeness should amplify rewards when suspicion is low or absent, and potentially elevate baseline rewarding even when signals are neutral. Our design models these predictions using an observer–centred approach (from DUTC report distributions) and exogenous variation in SES closeness/distance between players.

3.2.4 Hypotheses and contributions

Guided by the existing literature, we evaluate two hypotheses in a single framework:

- **H1: Punishment × Distance:** Observers punish suspicious reports more readily when the counterpart is socially distant.
- **H2: Reward × Closeness:** Observers reward non-suspicious reports more readily when the counterpart is socially close.

Because the suspicion signal is fixed by what P2 observed, differences by SES are interpretable as selective enforcement across social boundaries. Our paper advances the literature in three ways:

(1) A proxy to capture suspicious cheating at the Observer-level. We build a continuous, round-level proxy at the Observer-level. We flag statistically unusual spikes in specific reports within each Observer×SES cell (12 rounds), weight them by the marginal payoff of the reported die value (up-weighting payoff-maximising faces, down-weighting payoff-reducing ones) and attach the resulting score to that Observer's round. This retains the DUTC's aggregate inference, but shifts the signal down to the level of what each Observer actually sees in their respective treatment condition.

This proxy allows us to move beyond the traditional sample-level measure of cheating in the DUTC, while incorporating the strategic relevance of reports by weighting excess frequency with marginal payoffs. The ultimate objective is to propose a model producing a continuous signal varying across Observers, SES pairings and relevant rounds, while providing a continuous signal of suspicious cheating.

(2) A two-side, stake-free third-party punishment/reward framework. We analyse punishment and rewards as two sides of the same mechanism, mapping both onto a common scale and administering them by Observers whose own payoffs are unaffected. This design isolates responses to suspicious reports from motives linked to self-interest (e.g., retaliation, loss recovery, reciprocity).

(3) Socioeconomic selectivity. We manipulate socioeconomic closeness/distance between participants, objectively (by average income level in departments of residence) and subjectively (by personal income in comparison to departmental average), and test whether the same suspicion signal elicits different responses toward close vs. distant counterparts.

3.3 Experimental design and procedures

The experimental setting consists of three tasks: (1) a socioeconomic survey, (2) an online Die-under-the-cup (DUTC) task, and (3) the Moral Identity (MIS) scale. We ran the experiment online with three treatment conditions: Punishment (Treatment 1), Reward (Treatment 2) and Mixed (Treatment 3). In the following sections we provide a detailed overview of the experimental design and its tasks.

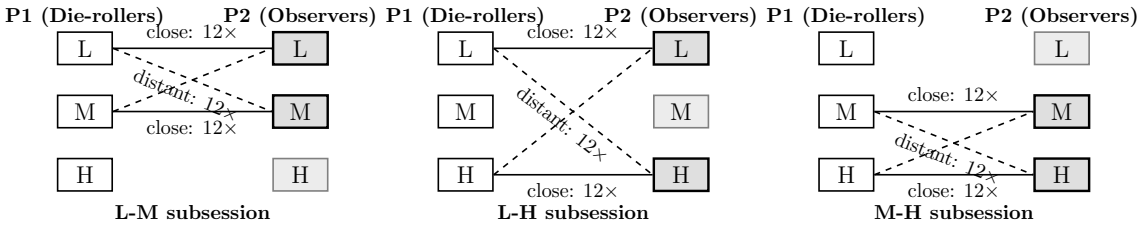
Participant recruitment and stratification

We ran the study online with participants from pool of voluntaries from the Laboratory of Experimental Economics of Paris (LEEP). At invitation, we pre-screened by postal code to map each

registrant to a department of residence and its corresponding SES tier (Low, Medium, High), using data from the French *National Institute of Statistics and Economic Studies (INSEE)*. Invitations were stratified by tier with tier-specific quotas to keep the pool balanced and ensure even pairings. Department coverage and tier assignment are reported in Appendix Table 3.31; a geographic visualisation appears in Figure 3.27.

Participant allocation followed a two-step plan: we first filled the *Mixed* treatment, as this was the largest and most complex condition ($n = 432$), then directed additional recruits to the *Punish* ($n = 144$) and *Reward* ($n = 144$) treatments. Participant roles (P1, P2) were randomised at login. Matching was organised in subsessions that included exactly two SES tiers (e.g., Low & Medium or Low & High), to ensure exact pairing cells within each subsession. For example, in a Low–Medium SES subsession, a Low-SES Observer (P2: Low) completed one 12-round block with 12 Low-SES Die-rollers (P1: Low; socially close) and one 12-round block with 12 Medium-SES Die-rollers (P1: Medium; socially distant). For a visual representation of SES-based pairs in the subsessions, see Figure 3.21.

Figure 3.21: Pairing by subsession (two-tier SES matching)



Solid line = same-tier (close pairings); dashed line = cross-tier (distant pairings). Shaded squares denote **P2 (Observers)**, the unit of analysis. Each P2 completes 12 close and 12 distant rounds per subsession.

Task 1. Socioeconomic questionnaire and participant categorisation

Participants completed a socioeconomic questionnaire covering demographics (age, gender, marital status, education, occupation) and socioeconomic information (personal monthly net income, partner's income, weekly working hours, housing status, credit situation and enrolment in social security programmes). For income-related questions, participants selected one of five ordered brackets ranging from 1 (under 1,000€) to 5 (over 6,000€), plus a non-applicable option. Given that a substantial share of participants were non-working students (and to a lesser extent unemployed individuals) who reported having no income, we use partner and household income as proxies.

Objective social distance

Objective SES is constructed from average monthly income at the departmental level. At the time of recruitment, we observe only participants' post codes (not personal income), therefore, we anchor SES tiers to INSEE's 2021 *niveau de vie médian* (median disposable income per

consumption unit) per month, at the department level INSEE Références (2021). Tiers are anchored to national reference points:

- **Low SES:** departmental median $<$ 1,993€ (national median);
- **Medium SES:** 1,993€ $<$ departmental median $<$ 2,477€;
- **High SES:** departmental median $>$ 2,477€ (top-income departments: Paris and Hauts-de-Seine).

This location-based SES captures the income environment of where participants live. Objective social distance is then defined at the pair level between Observer P_2 and Die-roller P_1:

$$\text{distance}_{P_2P_1} = \left|\text{SES}_P^{\text{obj}}{}_2 - \text{SES}_P^{\text{obj}}{}_1\right| \in \{0, 1\},$$

where 0 indicates same tier (socially close) and 1 different tiers (socially distant). For a comprehensive overview of the departments participants were recruited from and their categorisation by SES level, see Table 3.31.

Subjective social closeness

We construct a three–tier subjective SES from the income brackets reported in Task 1, mirroring the logic used for objective closeness. Let `player_revenue` be the participant's self-income bracket (1–5; 6 = no income). We define:

$$\text{SES}^{\text{subj}} = \begin{cases} \text{Low} & \text{if } \texttt{player\_revenue} \in \{1, 6\}, \\ \text{Medium} & \text{if } \texttt{player\_revenue} = 2, \\ \text{High} & \text{if } \texttt{player\_revenue} \in \{3, 4, 5\}. \end{cases}$$

We then run adjustments for participants with no personal income (`player_revenue` = 6), using partner income and, if needed, household income brackets:

$$\text{SES}^{\text{subj}} = \begin{cases} \text{High} & \text{if partner bracket} \in \{4, 5\}, \\ \text{High} & \text{if partner} = 6 \text{ and household} \in \{4, 5\}, \\ \text{Medium} & \text{if partner} = 3, \\ \text{Medium} & \text{if partner} = 6 \text{ and household} = 3, \\ \text{Low} & \text{otherwise.} \end{cases}$$

Finally, we compute the subjective closeness indicator by comparing it to the participant's objective SES tier from their department of residence ($\texttt{player\_tier} \in \{\text{Low}, \text{Medium}, \text{High}\}$):

$$\text{socially\_close\_subj} = \begin{cases} 1 & \text{if SES}^{\text{subj}} = \texttt{player\_tier}, \\ 0 & \text{if SES}^{\text{subj}} \neq \texttt{player\_tier}, \end{cases} \quad \text{and} \quad \text{socially\_distant}^{\text{subj}} = 1 - \text{socially\_close}^{\text{s}}$$

This procedure yields a clean subjective SES based primarily on personal income, with transparent partner/household reclassification only for respondents reporting no personal income, and a binary subjective closeness measure by matching subjective and objective tiers.

Socioeconomic status in experimental research

Modelling social closeness through socioeconomic status is grounded in both theoretical and empirical reasoning. A growing body of research identifies it as a powerful and multidimensional indicator of social identity and perceived interpersonal closeness (Easterbrook et al., 2019; Kraus et al., 2011; Manstead, 2018). Moreover, SES encompasses objective indicators, such as income, employment and education, as well as subjective dimensions, including perceptions, emotions, attitudes and moral values.

Using socioeconomic closeness as a proxy for social closeness allows us to analyse behavioural responses when cheating is committed by individuals from different socioeconomic backgrounds. In this sense, Kraus et al. (2011) found that income differences create strong in-group identity, particularly among working and middle-class individuals, which define their sense of belonging in opposition to out-groupers. This is further studied from a nationality-based perspective in Magni (2021), who show that greater inequality relates to greater favouritism towards nationals over immigrants in resource distribution. Similarly, Du et al. (2020) found that in societies with high perceptions of systemic income inequality, social norm enforcement tends to erode.

The literature presents divergent views regarding the effects of income inequalities for honesty: Chae et al. (2022) argue that resource scarcity increases in-group favouritism and induces group loyalty, hindering honesty. Rubin et al. (2014) show that individuals in lower income brackets tend to engage in in-group favouritism as a compensatory mechanism for their unfavourable position. Conversely, Elbæk et al. (2023) show that individuals with lower SES show stronger moral identity and prosocial intention, challenging the assumptions that low-income relates to a decline in honesty. This is supported in Martinangeli and Martinsson (2020), who reveal that when observing peer behaviour, low-income individuals become more norm-compliant when their in-group behaves honestly, whereas high-income individuals are more likely to justify dishonesty when their peers cheat.

We deem this analysis relevant in light of rising income disparities around the globe and the polarisation that societies increasingly experience. Our interest is to provide a novel perspective that links objective and subjective measures of closeness to suspicions of cheating in the DUTC.

Task 2. Online Die-under-the-cup (DUTC) task

Participants complete 24 consecutive rounds of a DUTC in fixed, randomly assigned roles:

- **Die-roller (P1):** privately rolls a die, observes the true outcome value, and registers a value for payment.
- **Observer (P2):** observes both the true outcome value and P1's registered value and, depending on the treatment, may apply a -2 penalty, a $+2$ reward, or simply observe; P2 then registers their own payoff.

Payoffs equal the registered die value in euros, except that a 6 yields 0. With a fair die each outcome occurs with probability 1/6, so under truthful reporting the expected mean is

$$\frac{1+2+3+4+5+0}{6} = 2.5.$$

One of the 24 rounds is randomly selected for payment at the end of the task.

Each round proceeds as follows:

1. **P1 private roll.** A virtual die appears on P1's screen. When P1 clicks *Roll*, the outcome is generated locally using the browser's cryptographic randomness; the true value $k \in \{1, \ldots, 6\}$ is shown to P1.

2. **P1 registration.** P1 enters a value to be used for payment.

3. **P2 observation.** P2 sees the true value and P1's registered value, applies the treatment-specific choice (penalise or reward), and then registers their own payoff for the round.

4. **Data handling.** After the round, the true outcome is not retained in the research dataset. Server logs store session metadata (session ID, round, treatment, role), registered values for payment, and P2's decision (punish/reward). True outcomes are not accessible to experimenters *ex post*.

Before each round, P2 is shown the counterpart's department-level average monthly income and its SES tier (Low/Medium/High), making socioeconomic distance salient without revealing identities.

Construction of the proxy for suspicion of cheating

Because experimenters do not observe P1's true outcome values, individual cheating cannot be directly identified after the experiment. We therefore construct a round-level suspicion signal that increases when two conditions jointly hold: (1) the value registered in the round is over-represented relative to a uniform benchmark and (2) that value is payoff-maximising for P1.

Setting the suspicion proxy.
Each Observer (P2) sees 24 registered values in total, split into two 12-round blocks: one with socially close counterparts and one with socially distant counterparts. Under truthful reporting, each face should appear with probability 1/6, so the expected count per face in a 12-round block is 2. Systematic counts above this benchmark indicate over-representation and are treated as suspicious. The proxy converts this abnormality into a round-level signal and weights it by payoff consequences (faces 4–5 are payoff-maximising; 1, 2, and 6 are payoff-reducing; 3 is near neutral). In what follows we detail step-by-step the proxy's construction:

Step 1: Count recorded faces within each block. For each Observer i, block $s \in \{\text{close}, \text{distant}\}$, and value $k \in \{1, \ldots, 6\}$, let

$$n_{ik}^{(s)} = \sum_{r=1}^{12} \mathbf{1}\{\text{recorded\_value}_{ir}^{(s)} = k\}$$

be the number of times value k was registered in the 12 rounds of block s.

Step 2: Compute excess frequency (over the uniform benchmark) Translate excess concentration into a unit-interval share:

$$b_{ik}^{(s)} = \begin{cases} \dfrac{n_{ik}^{(s)} - 2}{n_{ik}^{(s)}}, & \text{if } n_{ik}^{(s)} > 2, \\ 0, & \text{if } n_{ik}^{(s)} \leq 2. \end{cases}$$

Step 3: Attach the excess to the current round In round r of block s, let $k(r)$ denote the registered value. The round-level abnormality term is

$$E_{ir}^{(s)} = e_{i,\,k(r)}^{(s)}.$$

Step 4: Weighting by incentives (payoff alignment). Reports map into payoffs in euros (reports 1–5 pay 1–5€; reporting 6 pays 0€). To align the proxy with payoff-motivated misreporting, we weight by the marginal payoff consequences of each value:

$$S_{ir}^{(s)} = \sum_{k=1}^{6} w_k \, b_{ik}^{(s)} \, x_{ir}^{(s)}(k), \qquad (w_1, w_2, w_3, w_4, w_5, w_6) = (-1.5, -0.5, 0.5, 1.5, 2.5, -2.5).$$

Because only the reported value enters at round r, $S_{ir}^{(s)}$ reduces to the weight for that value times its excess share. Positive values indicate unusual concentration of payoff-maximising reports (4 or 5); negative values indicate unusual concentration of payoff-reducing reports (1, 2, or 6); values near zero are consistent with uniform play.

For instance, if in the 12 distant rounds for Observer i the value 5 appears 5 times, then

$$e_{i5}^{(\text{distant})} = \frac{5}{12} - \frac{1}{6} = 0.25,$$

so any distant round where 5 is registered has

$$S_{ir}^{(\text{distant})} = 2.5 \times 0.25 = 0.625 > 0.$$

Thereby reflecting a suspicious report in that particular round.
On the contrary, if in the 12 close rounds the value 1 appears 4 times, then

$$e_{i1}^{(\text{close})} = \frac{4}{12} - \frac{1}{6} = 0.1667,$$

and a close round registering 1 yields

$$S_{ir}^{(\text{close})} = -1.5 \times 0.1667 \approx -0.25.$$

Thereby reflecting an unsuspicious report in that particular round.

Treatment conditions

The experiment has three treatment conditions. At each round, Observers know: (1) the average monthly income in their department of residence, (2) the average monthly income in the Die-roller's department of residence, (3) the income tier of both departments, and (4) the national average monthly income in France, as a benchmark.

Die-rollers know that P2 can see both the true die outcome and the value they register for payment, and that P2 may have the ability to punish or reward depending on the assigned condition. However, they are not told the Observer's department of residence (proxy for social distance), the Observer's punish or reward decisions, nor their own round-level earnings until the experiment ends.

This serves two purposes: by withholding information about the Observer, we isolate Die-rollers' cheating behaviour from reputational concerns or reactive feedback, allowing for a cleaner measure of their suspicious cheating; and it models a realistic norm-enforcement scenario for Observers.

Treatment 1: Punishment

Player 1 (P1) rolls a die, obtains the true value $Y_{i,r}^{\text{true}} \in \{1, 2, 3, 4, 5, 6\}$ and reports a value $Y_{i,r}^{\text{rep}}$. Player 2 (P2) observes both $Y_{i,r}^{\text{true}}$ and $Y_{i,r}^{\text{rep}}$, and decides whether to apply a -2€ penalty from P1's payoff:

$$\text{Payoff}_{i,r} = \max(0, Y_{i,r}^{\text{rep}} - 2)$$

P2 then registers their own payoff $Y_{j,r}^{\text{rep}}$.
The Observer's decision to punish is independent of whether P1 cheated or not.

Treatment 2: Reward

P1 proceeds as in Treatment 1.
P2 observes both $Y_{i,r}^{\text{true}}$ and $Y_{i,r}^{\text{rep}}$, and decides whether to apply a +2€ reward to their payoff:

$$\text{Payoff}_{i,r} = \max(0, Y_{i,r}^{\text{rep}} + 2)$$

The Observer's decision to reward is also independent of whether P1 cheated or not.

Treatment 3: Mixed

P1 proceeds as in Treatment 1.
P2 is informed that there are three possible conditions they can be assigned to:

- **Punish:** Can apply a -2€ penalty to P1's payoff.
- **Reward:** Can apply a +2€ reward to P1's payoff.
- **Observer:** Only observes P1's behaviour.

At the start of the game, Observers are informed of the conditions they have been assigned to and that they will stay in this condition for the entirety of the game. They are also informed that this information will not be disclosed to P1.

Die-rollers are informed of the three possible conditions for Observers and that they will be randomly paired with 24 participants playing in one these roles. However, they are not informed of which type of Observer they are being paired with.

This treatment allows us to estimate anticipatory effects on whether the mere possibility of sanction or reward (as opposed to certainty) deters cheating or encourages honesty. Furthermore, it provides a cleaner test of social selectivity in enforcement, as any differences in P2's behaviour by social distance are less confounded by P1's strategic adaptation to a fixed regime.

Task 3. Moral Identity Scale (MIS)

We administered the Moral Identity Scale Aquino and Reed (2002) to capture stable individual differences in the centrality of moral traits to the self-concept. The instrument comprises two subscales—Internalisation (moral traits as part of one's self-view) and Symbolisation (expressing moral traits outwardly)—with items rated on a 1–10 scale (1 = "not at all," 10 = "very much"). Traits such as *honest*, *fair* and reverse-coded antonyms (e.g., *ruthless*, *selfish*) were included. The MIS was administered after the DUTC to avoid priming moral concerns during enforcement decisions.

Because MIS captures broad, trait-like heterogeneity that can absorb meaningful between-subject variation in enforcement, including it as a main regression control risks over-controlling and complicating interpretation of treatment and closeness effects. We therefore decided to exclude this variable from our analyses and from main regressions.

Experimental procedures

The experiment was run online in July of 2025. We recruited 780 volunteers, after applying pre-specified exclusions for incomplete participation, we excluded 60 participants, yielding a final sample of $n = 720$ participants (60% female), with a mean age of 28.6 years (SD = 9.6). The pool consisted mainly of students (41.5%) and employed individuals (39%); see Appendix Tables 3.27 and 3.28 for education and occupation details. Average earnings were €6.14 (SD = €4.88), including a €3 show-up fee.

Participants were assigned at login to one of three session-level treatments: *Punish* ($n = 144$), *Reward* ($n = 144$), or *Mixed* ($n = 432$). Each participant completed 24 rounds, for a total of 17,280 round-level observations: 3,456 in *Punish* (144×24), 3,456 in *Reward* (144×24), and 10,368 in *Mixed* (432×24).

Round-level observations are not independent because each participant contributes 24 decisions. Accordingly, all inference uses cluster-robust standard errors at the participant level. For P2 decisions, standard errors are clustered by *Observer*; for models of P1 reporting, they are clustered by *Die-roller*. As a robustness check, we also report specifications with two-way clustering by P1 and P2 when both actors' histories may induce dependence. Regression tables report the total number of round-level observations and the number of clusters used for inference.

3.4 Descriptive statistics

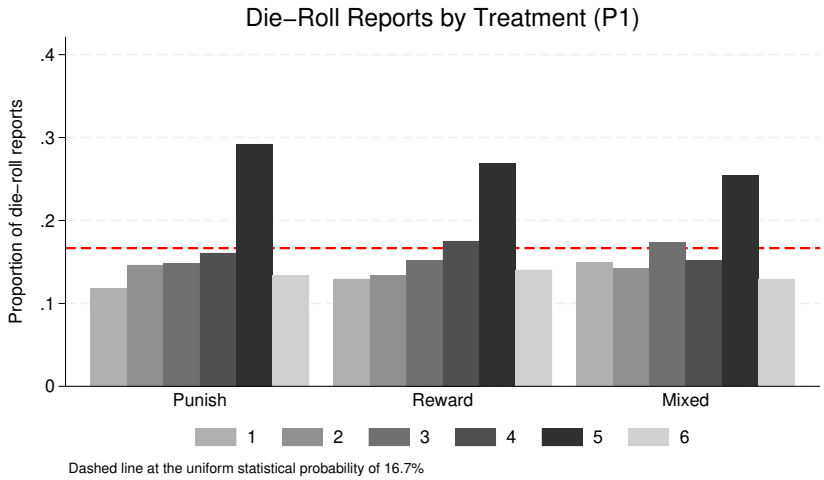
3.4.1 Die-roll reports and potential cheating propensities

Figure 3.22 shows the distribution of Die-rollers' reported outcomes by treatment. As in standard DUTC studies, reports of 5 are over-represented relative to the uniform benchmark of 16.7%, reports of 1, 2, and 6 are under-represented, and reports of 3 and 4 are closer to uniform. For inference, we collapsed to participant–treatment shares (one observation per P1×Treatment) and ran OLS of the share of 5s on treatment dummies with standard errors clustered by participant. This yields $N = 360$ cells (Punish: 72, Reward: 72, Mixed: 216). A joint test of equality of the

three treatment means (cluster-robust Wald) yields $F(2,;\mathrm{df}_{\mathrm{cluster}}) = 0.70$, $p = 0.497$.[1] Pairwise differences are: Reward − Punish $\Delta = -0.023$, $p = 0.565$; Mixed − Punish $\Delta = -0.038$, $p = 0.245$; Reward − Mixed $\Delta = 0.014$, $p = 0.657$.

Within each treatment, we tested the mean share of 5s against 1/6 using one-sample regressions with participant-clustered errors. Estimated means confirm excess of payoff-maximising reports in all treatments: Punish $\hat{\mu} = 0.292$ (72 clusters; $F(1, 71) = 18.39$, $p = 0.000$); Reward $\hat{\mu} = 0.269$ (72; $F(1, 71) = 12.28$, $p = 0.000$); Mixed $\hat{\mu} = 0.254$ (216; $F(1, 215) = 34.54$, $p < 0.001$).

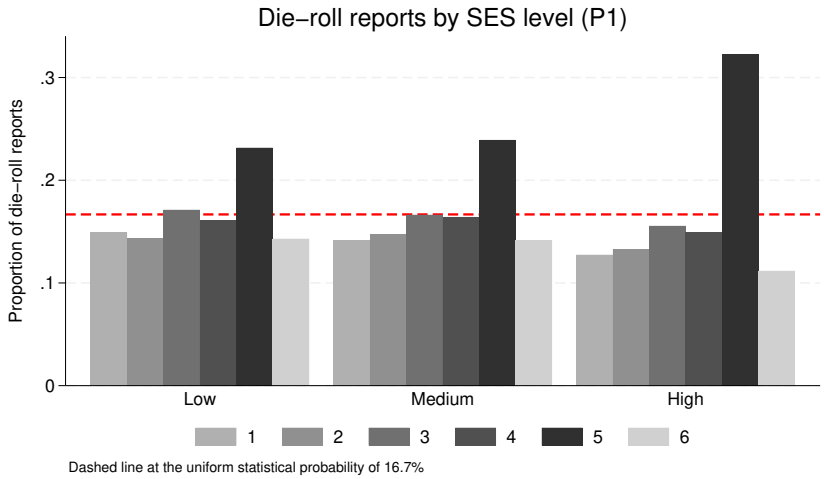
Figure 3.22: Die-roll reports by treatments (P1)



Total number of observations: N = 17,280. By treatment: Punish (N = 3,456), Reward (N = 3,456), Mixed (N = 10,368). The figure displays round-level proportions; participant–treatment means (one observation per P1×Treatment, N = 360) are reported in the text. Mean round-level proportions of reported outcomes 4 and 5 are: Punish — 4 = 0.157, 5 = 0.354; Reward — 4 = 0.190, 5 = 0.304; Mixed — 4 = 0.155, 5 = 0.326.

Figure 3.23 disaggregates P1 reports by SES, showing that trends persist: excess reports of 5, fewer reports of 1, 2 and 6. For inference, we collapsed to participant-level shares (one P1×SES observation per participant) and regressed the share of 5s on SES indicators with standard errors clustered by participant (N = 8,640 round-level observations mapped into 360 clusters). A joint test across tiers gives $F(2, 359) = 4.89$, $p = 0.008$. Relative to High-SES, Low-SES ($\Delta = -0.091$, $p = 0.003$) and Medium-SES ($\Delta = -0.083$, $p = 0.009$) report fewer 5s. One-sample clustered tests against 1/6 show that each tier still exceeds the uniform benchmark (Low: $F(1, 119) = 37.67$, $p < 0.001$; Medium: $F(1, 119) = 15.47$, $p < 0.001$; High: $F(1, 119) = 14.23$, $p < 0.001$). Moreover, an ordered trend (High= 1, Medium= 2, Low= 3) indicates a decline in reports of 5 as SES decreases: $\hat{\beta} = 0.0457$ (two-sided $p = 0.003$). In short, the spike in payoff-maximising reports is present for all SES groups and is strongest among High-SES participants (see also Alfonso-Costillo et al. (2022); Fischbacher and Föllmi-Heusi (2013); Hermann and Ostermaier (2018); Kocher et al. (2018)).

Figure 3.23: Die-roll reports by SES (P1)

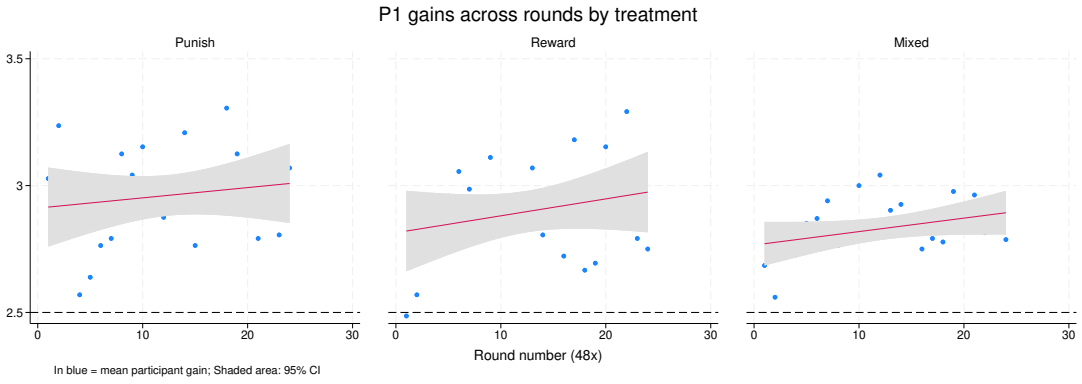


Total: N = 8,640. By SES level: Low (1) = 2,880, Medium (2) = 2,880, High (3) = 2,880.
Mean frequency of die-roll reports (by P1): Low = 0.393, Medium = 0.403, High = 0.472.

Figure 3.24 plots round-by-round mean P1 gains by treatment, with a reference line at the truthful benchmark (2.5). Mean gains exceed 2.5 in all treatments (Punish = 2.962, Reward = 2.898, Mixed = 2.832; each $p < 0.001$ from clustered one-sample tests), consistent with excess in payoff-maximising reports. Using round-level data with participant-clustered errors, mean gains do not differ across treatments ($F(2, 359) = 0.70$, $p = 0.500$), nor do trends diverge systematically over rounds. This suggests treatment differences—if any—operate via the composition of reported values rather than sustained shifts in average earnings.

Because true P1 outcomes are anonymous to researchers, suspicion is inferred by combining these gain patterns with report distributions and P2 decisions. Intuitively, if P2s persistently penalised high reports, mean gains would fall below 2.5 in Punish; if P2s persistently rewarded moderate reports, mean gains would sit above 2.5 in Reward. The similarity across panels indicates no net punishment or reward effect on average earnings over time. The high similarity in both graphs shows no evidence of a net punishment or net reward effect on average P1 earnings across rounds.

Figure 3.24: P1 mean earnings by treatment across 24 rounds

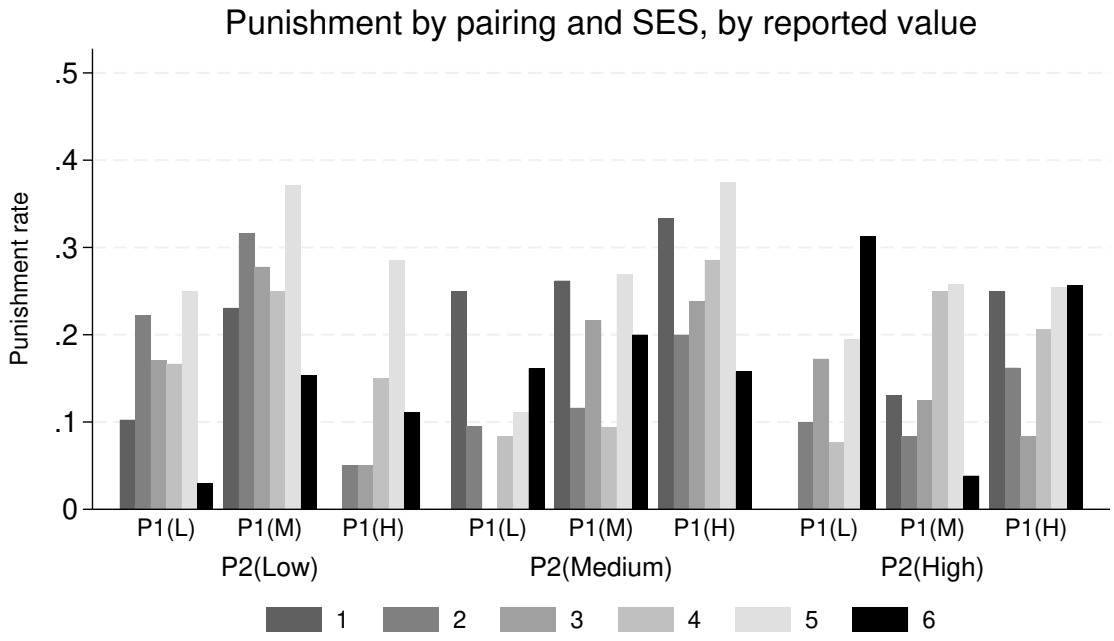


3.4.2 Attitudes towards cheating: punishment vs reward

Figure 3.25 reports mean punishment rates by P2 SES × P1 SES × reported values. Punishment peaks at reports of 5 across pairings, with reports of 6 also attracting non-trivial punishment despite paying 0€. We relate to either a "larger number = more cheating" heuristic among some Observers, or to occasional misreadings of the payoff rules.

Conditioning on the same report, SES modulates severity more than scope. For instance, at 5, P2:High punish P1:Low/Med/High at comparable rates (0.194/0.258/0.254). The starkest asymmetry appears at 6, where P2:High are harsher toward P1:Low (0.312) relative to P2:Low→P1:Low (0.029) and P2:Med→P1:Low (0.161). On average, P2:Medium are the most punitive, while P2:Low the least. Importantly, same-SES cells (L-L, M-M, H-H) do not exhibit uniform leniency, as their rates at 5 cluster around one quarter, and at 6 range from very low (L-L) to mid-range (H-H). We also note that a few cells display zero observed punishment (e.g., P2:Low×P1:High at 1), which reflect small-N cells rather than missing data, therefore we annotate the figure so zero-height bars are visible rather than mistaken for missing observations. Appendix Table 3.32 reports, for every P2×P1×report cell, the counts and rates with exact confidence intervals.

Figure 3.25: Propensity to punish

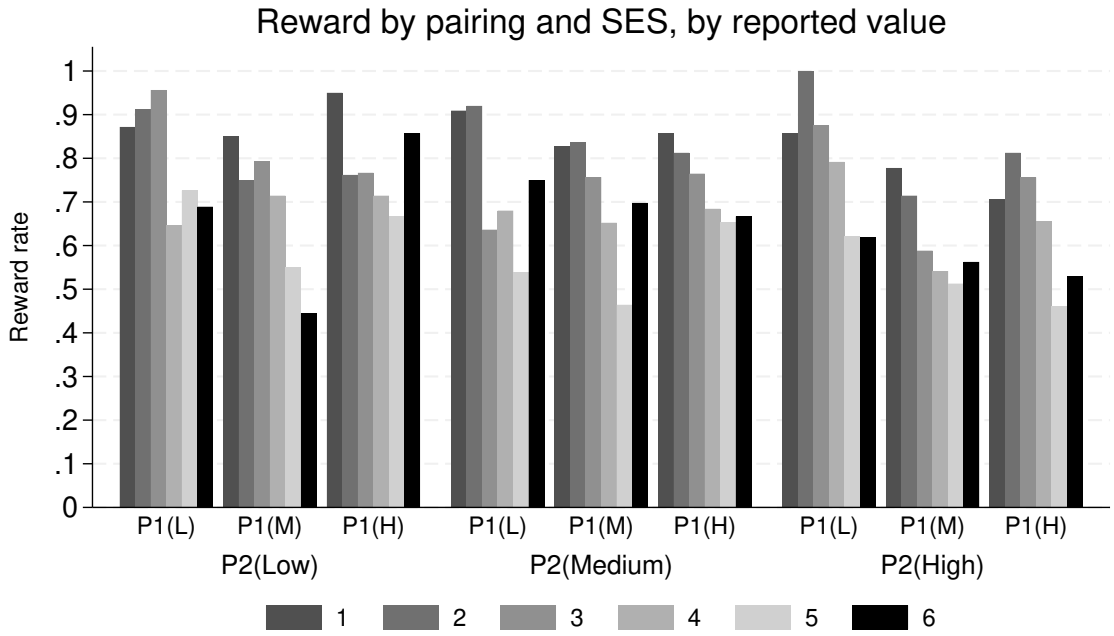


Graph reports on P2 decisions to punish P1 ($N = 712$) out of all instances where punishing was possible (Punish treatment + Punish condition of Mixed treatment, $N = 3{,}456$). Ranges in number of observations correspond to different reported die values being rewarded: P2:H rewarded P1:H counterparts N=34–90 (49–70%), P1:M N=14–34 (53–80%), and P1:L N=14–51 (55–93%). P2:M rewarded P1:H N=12–52 (22–70%), P1:M N=16–65 (67–79%), and P1:L N=16–46 (55–78%). P2:L rewarded P1:H N=14–43 (64–81%), P1:M N=16–49 (59–83%), and P1:L N=39–73 (69–94%).

Figure 3.26 shows the mirror pattern for rewards. Across SES pairings, reward rates are generally high at reports 1-3 and drop at reports 5 and 6. For example, P2:Low reward P1:Low at 0.872, 0.912, and 0.957 for reports 1-3, and P2:High reward P1:Low at 0.857, 1.000, and

0.875 over the same range. By contrast, at report 5 several pairings exhibit the lowest rewards in their row, such as P2:Medium×P1:Medium at 0.464 and P2:High×P1:High at 0.461, while some pairings remain comparatively generous (e.g., P2:Low×P1:Low at 0.726). Report 6 is heterogeneous rather than uniformly low, with high rewards for P2:Low×P1:High (0.857) but lower values for P2:High×P1:High (0.529) and P2:High×P1:Med (0.562). Full counts and 95% CIs are in Appendix Table 3.33.

Figure 3.26: Propensity to reward



Graph reports on P2 decisions to reward P1 (N = 2,397) out of all instances where rewarding could have occurred (Reward treatment + reward condition of Mixed treatment, N = 3,456). Since punishment was not implemented, P2 never punished any P1 counterpart; the number of observations varies by reported die value: P2:H counterparts P1:H N=34–89 per die value, P1:M N=14–39, and P1:L N=14–37. P2:M counterparts P1:H N=12–52, P1:M N=16–49, and P1:L N=18–40. P2:L counterparts P1:H N=14–35, P1:M N=16–46, and P1:L N=39–73.

3.5 Results

In this section we reports the results from our sample analyses. Unless stated otherwise, we estimate logistic regressions and report average marginal effects (AMEs) with cluster-robust standard errors by Observer (P2). Since each P2 makes 24 decisions, round-level observations are dependent, therefore, all models include round fixed effects to absorb common time patterns across the 24 rounds.

We address unobserved, stable differences across Observers in two ways. First, we estimate conditional logit models with Observer fixed effects, which use only within-Observer variation. Second, as a robustness check, we estimate linear probability models (LPM) with Observer fixed effects. These yield the same qualitative conclusions. For punishment (reward) regressions, we use rounds from the Punish (Reward) treatment and, from Mixed, only Observers assigned the corresponding condition. Observe-only rounds are excluded from enforcement regressions.

3.5.1 Decisions to punish across objective social distance

Hypothesis 1. Observers' propensity to punish increases when the counterpart is socially distant (vs. close).

The dependent variable is `punish`, equal to 1 when P2 applies a -2 penalty to P1's payoff in rounds where punishment is enabled (i.e., the Punish treatment; the punishment condition of the Mixed treatment), and 0 otherwise. We use three regressors:

- **Suspicion of cheating:** `suspicious_report`, a round-level index that rises with unusually frequent payoff-maximising reports in the relevant Observer×SES exposure.
- **Objective social closeness:** `socially_close`, 1 if P1 and P2 reside in departments in the same SES tier, 0 otherwise.
- **Punishing close counterparts:** `suspicious_report_close`, interaction of previous variables, capturing whether responsiveness to the suspicion signal differs for socially close pairs.

The estimation sample pools all rounds from the Punish treatment and rounds from T3 (Mixed) for Observers assigned the Punish conditions.

Results show that punishment is responsive to the suspicion signal:(`suspicious_report` = 0.244, $p < 0.010$). Social closeness reduces this responsiveness: (`suspicious_report_close` = -0.164, $p < 0.10$), so the slope for socially close pairs is smaller ($0.244 - 0.164 \approx 0.08$). Holding the signal fixed, socially close counterparts are punished less: (`socially_close` = -0.326, $p < 0.05$).

Disaggregating analyses by counterpart SES (Low as reference), distant pairs punish Medium and High more leniently than Low: (`SES Medium` = -0.352, $p < 0.050$) and (`SES High` = -0.424, $p < 0.010$). Social closeness offsets this pattern for High-SES counterparts (`High×close` = 0.561, $p < 0.01$), while the Medium interaction is small and not significant (`Medium×close` = 0.0887, $p = 0.705$).

Among controls, punishment is higher among women ($+0.227$, $p < 0.050$) and increases with age ($+0.021\, p < 0.010$); it is unrelated to occupation ($+0.039$, $p = 0.171$) and lower with education (-0.190, $p < 0.010$).

In sum, punishment rises with stronger suspicion, but this responsiveness is attenuated for socially close pairs. Conditional on the same suspicion level, Low-SES counterparts are more punished than Medium or High-SES ones. Although P1:High report 5s more often on average (Figure 3.23), Observers still sanction P1:Low more at a given suspicion level, consistent with socioeconomic selectivity in enforcement. The apparent leniency toward High-status counterparts is mostly a feature of distant pairings and disappears when counterparts are close.

Table 3.23 reports AMEs from the main logit with SEs clustered by Observer; robustness with two-way clustering (P2 and P1), Observer fixed effects (conditional logit), and LPM with Observer fixed effects yields qualitatively identical conclusions.

Table 3.23: Punishment of socially close counterparts (objective)

Variables	Punishment
suspicious_report	0.244***
	(0.058)
suspicious_report_close	-0.164*
	(0.084)
SES Medium	-0.352**
	(0.147)
SES High	-0.424***
	(0.148)
socially_close	-0.326**
	(0.150)
SES medium#Socially_close	0.088
	(0.211)
SES high#Socially_close	0.561***
	(0.210)
Gender (female = 1)	0.227**
	(0.092)
Age	0.0210***
	(0.004)
Occupation	0.039
	(0.027)
Scholarity	-0.190***
	(0.041)
Constant	0.603
	(0.439)
Observations	3,456

Robust standard errors in parentheses.$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.5.2 Decisions to punish across subjective social distance

We re-estimate the punishment model replacing objective closeness with the subjective measure. The dependent variable is `punish` (1 if P2 applies a -2 penalty in rounds where punishment is enabled; 0 otherwise). Key regressors:

- **Suspicion of cheating:** `suspicious_report`.
- **Subjective closeness:** `socially_close_subjective` =1 if P2's self-reported income tier matches P1's departmental tier; 0 otherwise.
- **Suspicion × subjective closeness:** `suspicious_report_subjective` (interaction).

Controls include SES dummies for the counterpart (Medium, High; Low is the reference), their interactions with `socially_close_subjective`, and controls. We estimate logit models, report average marginal effects (AMEs), use round fixed effects, and cluster standard errors by Observer (P2). Robustness checks with two-way clustering (P2&P1), conditional logit (P2 FE), and LPM with P2 FE yield similar conclusions.

Punishment responds positively to the suspicion signal (`suspicious_report` = 0.245, $p <$ 0.010). Unlike the objective measure, moderation by subjective closeness is not significant

Table 3.24: Punishment of socially close counterparts (subjective)

Variables	**Punishment**
suspicious_report	0.245***
	(0.054)
suspicious_report_subjective	-0.127
	(0.085)
SES Medium	-0.357***
	(0.135)
SES High	-0.329**
	(0.141)
Socially close subjective	0.168
	(0.162)
SES Medium#Socially close subjective	0.236
	(0.240)
SES High#Socially close subjective	0.434*
	(0.233)
Gender (female = 1)	0.287***
	(0.096)
Age	0.020***
	(0.004)
Occupation	0.0495*
	(0.029)
Scholarity	-0.202***
	(0.042)
Constant	0.38
	(0.440)
Observations	3,456

Robust standard errors in parentheses $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

(`suspicious_report_subj` $= -0.127$, $p = 0.138$), and there is no baseline difference by subjective closeness when holding the signal fixed (`socially_close_subjective` $= 0.168$, $p = 0.300$).

By counterpart SES (vs. Low), punishment is lower for Medium and High (`SES Medium` $= -0.357$, $p < 0.010$; `SES High` $= -0.329$, $p < 0.05$). Interactions with subjective closeness are not significant for Medium ($+0.236$, $p = 0.325$) and only marginal for High-SES ($+0.434$, $p < 0.10$). Moreover, punishment is higher among women ($+0.287$, $p < 0.010$) and with age ($+0.020$, $p < 0.010$), weakly higher with occupation ($+0.0495$, $p < 0.100$), and lower with education (-0.202, $p < 0.010$).

Consistent with H1, punishment increases with the suspicion index. However, using the subjective closeness measure we do not detect a reliable distance effect (neither a main effect of closeness nor a stable attenuation of the suspicion–punishment slope). The SES pattern persists: at comparable suspicion, Medium and High-SES counterparts are punished less than Low-SES counterparts, with only a marginal uptick when High-SES targets are subjectively close. Overall, these results support the evidence-responsiveness component of H1 but do not provide clear support for a subjective-closeness moderation. See Table 3.24) for full estimates (AMEs), clustered SEs by P2, and listed robustness checks.

3.5.3 Decisions to reward across objective social distance

Hypothesis 2. Observers' propensity to reward increases when the counterpart is socially close.

The dependent variable is `reward` (1 if P2 adds +2 to P1's payoff in reward-enabled rounds; 0 otherwise). We estimate a logit and report average marginal effects with standard errors clustered by Observer (P2); round fixed effects are included. The estimating sample pools the Reward treatment and, from Mixed, only Observers assigned reward conditions (Observe-only rounds excluded). We use three key regressors:

- **Suspicion of cheating:** `suspicion_gain`.
- **Objective social closeness:** `socially_close` P1 and P2 in the same SES tier.
- **Moderation:** `suspicion_gain_close` = `suspicion_gain` × `socially_close`.

And add `ses` tier dummies for P1 (reference: Low-SES), their interactions with `close`, and demographic controls.

Rewards move inversely with the suspicion index (`suspicion_gain` $= -0.241$, $p < 0.010$), indicating more rewards when reports look non-suspicious. Objective closeness does not change this responsiveness (`suspicion_gain_close` $= 0.008$, $p = 0.908$). Holding suspicion fixed, socially close counterparts receive more rewards (`socially_close` $= 0.578$, $p < 0.010$).

By counterpart SES, main effects are small and not significant (Medium $= -0.106$, $p = 0.409$; High $= 0.193$, $p = 0.147$), but closeness effects attenuate with status: Medium×Close $= -0.531$, $p < 0.01$; High×Close $= -0.977$, $p < 0.01$. Moreover, women reward less (-0.166, $p < 0.050$), while age and occupation are not significant (-0.003, $p = 0.415$; -0.027, $p = 0.237$); education is positively associated with rewarding (0.092, $p < 0.010$).

In sum, results show that rewards are higher when suspicious cues are weaker. However, closeness does not alter sensitivity to the suspicion signal, but it does shape who benefits at a given level of suspicion: close, low-SES counterparts are favoured, the advantage fades for Medium and reverses for High-SES. Robustness (two-way clustering by P1 and P2; conditional logit with Observer FE; LPM with Observer FE) yields the same significant conclusions.

Table 3.25: Rewards of socially close counterparts (objective)

Variables	Reward
suspicious_report	-0.241***
	(0.051)
suspicious_report_close	0.0084
	(0.073)
SES Medium	-0.106
	(0.129)
SES High	0.193
	(0.133)
Social closeness	0.578***
	(0.141)
SES Medium#Socially close	-0.531***
	(0.187)
SES High#Social closeness	-0.977***
	(0.188)
Gender (female = 1)	-0.166**
	(0.080)
Age	-0.003
	(0.003)
Occupation	-0.027
	(0.023)
Scholarity	0.092***
	(0.033)
Constant	0.535
	(0.385)
Observations	3,456

Robust standard errors in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.5.4 Decisions to reward across subjective social distance

We re-estimate the reward model replacing objective closeness with the subjective measure. The dependent variable is `reward` (1 if P2 adds +2 in reward-enabled rounds; 0 otherwise). We estimate a logit and report average marginal effects (AMEs) with standard errors clustered by P2. We use the following key regressors:

- **Suspicion of cheating:** `suspicion_gain`.
- **Subjective social closeness:** `socially_close_subjective` (P2's subjective SES matches P1's departmental SES tier).
- **Moderation:** `suspicion_gain_subjective` = `suspicion_gain` × `socially_close_subj`.

Rewards move inversely with the suspicion index (`suspicion_gain` $= -0.313$, $p < 0.01$), indicating more rewards when reports look non-suspicious. Unlike the objective specification, subjective closeness does moderate this responsiveness: `suspicion_gain_subjective` $= 0.210$ ($p < 0.010$). Given the negative baseline slope, this implies a substantially weaker decline in rewarding as suspicion rises for subjectively close pairs (about $-0.313 + 0.210 \approx -0.103$). Holding other regressors fixed, there is no baseline difference by subjective closeness (`socially_close_subjective` $= -0.088$, $p = 0.575$).

By counterpart SES, rewards are lower for P1:High (Medium = −0.460, $p < 0.010$; High = −0.393, $p < 0.010$). Subjective closeness partly offsets this for P1:Medium (Medium×Close = 0.385, $p = 0.070$) and P1:High (High×Close = 0.438, $p = 0.060$), both with marginal numbers. Among controls, education is positively associated with rewarding (0.075, $p < 0.050$), while gender, age and occupation are not significant at conventional levels.

Consistent with H2's evidence channel, Observers reward less as suspicious cues grow stronger. Under the *subjective* closeness measure, this sensitivity is attenuated for subjectively close pairs: when P2 feels close to P1, the decline in rewarding with rising suspicion is smaller. Baseline rewarding does not differ by subjective closeness at a given suspicion level. Status patterns persist—Medium- and High-SES Rollers are rewarded less than Low-SES at comparable suspicion—though subjective closeness marginally tempers this gap.

Rewards move inversely with the suspicion index: observers are more likely to reward when reports appear non-suspicious. Unlike the objective measure, subjective closeness moderates responsiveness, as the decrease in rewards when suspicions of cheating rise is weaker for subjectively close pairs. Moreover, there is no baseline difference between subjectively close and distant counterparts at a given suspicion level. By counterpart SES, rewards are lower for medium and high-SES than for low-SES at comparable suspicion levels, while subjective closeness partly offsets this for medium and high status but only marginally. In short, under the subjective measure of closeness, as suspicion of cheating goes up, most Observers reward less. But when they feel subjectively close to the person, that drop in rewarding is smaller.

Table 3.26: Rewards of socially close counterparts (subjective)

Variables	**Reward**
suspicious_report	-0.313***
	(0.044)
suspicious_report_subjective	0.210***
	(0.079)
SES Medium	-0.460***
	(0.118)
SES High	-0.393***
	(0.117)
Socially close subjective	-0.088
	(0.158)
SES Medium#Socially close subjective	0.385*
	(0.213)
SES High#Socially close subjective	0.438*
	(0.233)
Gender (female = 1)	-0.112
	(0.081)
Age	-0.003
	(0.004)
Occupation	-0.002
	(0.024)
Scholarity	0.075**
	(0.034)
Constant	0.663*
	(0.395)
Observations	3,456

Robust standard errors in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.6 Concluding Remarks

This paper investigated whether third–party norm enforcement in a Die–under–the–Cup setting is associated with social closeness and socioeconomic status. Across four regressions that introduced decisions to punish and reward, each with an objective and a subjective measure of closeness, the main result is clear: enforcement is predominantly driven by suspicions of cheating and asymmetrically associated with socioeconomic status. Our findings show that decisions to punish increase in parallel to suspicions of cheating, while decision to reward increase as these same suspicions weaken.

A key component of our study is the proposed proxy to capture suspicious reports in the DUTC, as this allowed us to estimate responses that faithfully depict norm enforcement. The fact that punishment rises and reward falls along with the index across specifications supports the proxy's construct validity and portability to other private–report contexts, especially when anonymity needs to be preserved in order to obtain more reliable answers.

As for the selectiveness of norm enforcement, our results provide a clear picture of how social closeness shapes the baseline level of sanctions and, in some cases, how sensitive those sanctions are to the suspicion signal. Under our objective measure of closeness, close counterparts are punished less than distant ones at the same suspicion level. For rewards, close pairs have a higher baseline likelihood of being rewarded at a given suspicion. This baseline difference is largest for low-SES counterparts, small for medium-SES, and reverses for high-SES. Moreover, the reward–suspicion slope does not change with objective closeness, so the difference comes from levels rather than responsiveness. Under subjective closeness, punishment levels and slopes remain tied to the signal with no reliable changes, while rewards show reduced sensitivity to suspicion: as suspicion rises, rewards decline more slowly for subjectively close pairs, without a systematic lift or drop in baseline levels. Our results not only illustrate the way in which capturing closeness from different angles can shift norm enforcement, but also that a simple mechanism such as living in the same area can induce feelings of closeness.

On the other hand, socioeconomic status interacts with these patterns in a systematic way. One of the most telling results is that socioeconomic status interacts with these patterns to punish low–SES counterparts more than higher SES ones for the same suspicion signal, echoing previous studies suggesting that socioeconomic bias is negatively skewed against lower-income individuals. At the same time, closeness to high-status counterparts removes much of the leniency observed elsewhere, implying that the more punitive observers to high-income participants are other high-income individuals. This signal higher standards among high SES participants, reducing the gap in treatment. On the reward side, the asymmetry is complementary: under objective closeness, extra rewards concentrate on low-status counterparts, are small for medium-status, and turn negative for high-status. Under subjective closeness, the baseline premium fades, but closeness softens the decline in rewards as suspicion grows, indicating a closeness-based tolerance to ambiguous signals rather than a general lift in generosity.

Overall, our findings speak to the literature on selective norm enforcement and status expectations. They show that behaviour drives sanctions, with distance and status selectively tilting how strongly observability is converted into action. They also clarify why studies alternately find distance–based asymmetries or null effects: objective, place–based closeness shifts baselines

and interacts with status; subjective, *post hoc* similarity softens the link between evidence and generosity without moving baselines.

Limitations and future directions There are two caveats that frame interpretation in our analyses. First, while the socioeconomic status is consistent with expectation–based standards, isolating mechanisms such as inequality aversion or deservingness would require a design that orthogonally vary inequality transparency and attribution cues to further infer the impact of SES differences. Second, while modelling socioeconomic closeness at the department level proves a reliable measure, it may blur heterogeneity within large or unequal areas in more fine-grained analyses. Finer geographic income data could sharpen identification. Future work should (1) test other distance dimensions (friendship, politics, ethnicity, workplace), (2) examine contexts where Observers bear costs or face accountability for sanctions, (3) and replicate in field or organisational settings with richer administrative data.

Policy implications There are two leading implications for practical implementation of our analyses. First, decisions are evidence-driven: rewards rise when suspicion is low and punishments rise when suspicion is high. Second, enforcement is socially selective: leniency for close others and harsher sanctions for low-status counterparts at comparable evidence levels. Organisations can act on both: make evidence the focal point (clear thresholds, structured checklists), and reduce identity-based skew (mask socio-economic cues where possible; separate evidence review from identity; use blind panels). Reward mechanisms that recognise clear "non-suspicion" can strengthen compliance without stigmatising low-status groups; conversely, sanction policies should be audited for SES-linked disparities and calibrated with transparent guidelines.

Contribution and originality. We recast third-party enforcement as an evidence-to-action process and show how social closeness shapes that conversion. Methodologically, we build a continuous suspicion index anchored in chance and incentives, using only what each observer actually sees. This observer-level, payoff-aligned signal preserves the DUTC's core logic while moving beyond population averages to the decision environment each observer faced. Our design is stake-free and two-sided: the same informational signal applies to both, punishment and reward enforcement, allowing clean comparisons of how identical evidence translates into sanctions versus bonuses.

Substantively, the structure lets us separate who the target is from how evidence is used. Moreover, by manipulating closeness with two measures, we can provide inferences on how objective closeness mainly shifts levels of enforcement, tilting the average treatment people receive, while subjective closeness mainly shifts responsiveness, altering how sharply observers react to the same evidence.

Conditioning on the suspicion index makes status-linked selectivity salient to observe how socioeconomic differences change approaches to enforcement. These patterns refine accounts of in-group favouritism by showing that proximity does not replace evidence, but it biases the conversion of the same evidence into action. Furthermore, the suspicion index is computed only from the distribution of registered values that an Observer sees and the payoff mapping, so it can be applied wherever truthful answers or reporting is hidden by design. Because punishment and reward are both responses to the same signal and carry no stakes for Observers, we obtain a unified measurement of enforcement that is directly comparable across levers.

3.7 Appendix

Table 3.27: Demographics on scholarity

Diploma	**Observations**	**Participants**
PhD (8+ years)	1,104	46
Master (5 years)	8,664	361
Bachelor (3–4 years)	4,248	177
Technical degree (2 years)	1,440	60
High school diploma	1,536	64
Secondary education	192	8
No diploma	96	4
Total	17,280	720

Table 3.28: Demographics on occupation

Occupation	**Observations**	**Participants**
Employed (manager or independent)	2,928	122
Employed	3,840	160
Independent	936	39
Pensioner	312	13
Unemployed	1,704	71
Student	7,128	297
Working students	3,768	157
Other	432	18
Total	17,280	720

Table 3.29: Distribution of die-roll reports by Die-rollers (P1)

Roll value	**Observations**	**Observed %**	**Expectation**	**Δ**
1	1,204	13.9%	≈ 16.7%	-2.8
2	1,222	14.1%	≈ 16.7%	-2.6
3	1,420	16.4%	≈ 16.7%	-0.3
4	1,367	15.8%	≈ 16.7%	-0.9
5	2,286	26.5%	≈ 16.7%	+9.8
6	1,141	13.2%	≈ 16.7%	-3.5

16.7% is the expected uniform distribution for each reported die-roll; Δ is the difference from each aggregate reported outcome to the expected value in percentage.

Table 3.30: Distribution of die-roll reports by Observers (P2)

Roll value	**Observations**	**Observed %**	**Expectation**	**Δ**
1	725	8.39%	≈ 16.7%	-8.31
2	993	11.49%	≈ 16.7%	-5.21
3	1,306	15.12%	≈ 16.7%	-1.58
4	1,434	16.60%	≈ 16.7%	-0.10
5	3,371	39.02%	≈ 16.7%	+22.32
6	811	9.39%	≈ 16.7%	-7.31

Table 3.31: Departments of residence in the sample (by observations)

Department	**Observations**	**Percentage (%)**
Bas-Rhin	1,656	9.58
Bouches-du-Rhône	264	1.53
Essonne	552	3.19
Haut-Rhin	192	1.11
Hauts-de-Seine	1,656	9.58
Maine-et-Loire	120	0.69
Meurthe-et-Moselle	240	1.39
Moselle	600	3.47
Nord	432	2.50
Paris	4,104	23.76
Pas-de-Calais	192	1.11
Rhône	144	0.83
Seine-Maritime	552	3.19
Seine-Saint-Denis	1,968	11.39
Seine-et-Marne	312	1.81
Val-d'Oise	504	2.92
Val-de-Marne	648	3.75
Yvelines	384	2.22
Other	2,760	15.97
Total	17,280	100.00

Only departments with more than 100 observations (i.e., at least 5 participants) are shown in the table. The rest are compiled under "Other".

Table 3.32: Punishment by P2 SES × P1 SES and reported value: rate [95% CI]

P2-P1 pairing	Report 1	Report 2	Report 3
P2 Low – P1 Low	0.103 [0.029, 0.242] (39)	0.222 [0.120, 0.356] (54)	0.171 [0.072, 0.321] (41)
P2 Low – P1 Med	0.231 [0.050, 0.538] (13)	0.316 [0.126, 0.566] (19)	0.278 [0.142, 0.452] (36)
P2 Low – P1 High	0.000 [0.000, 0.308] (10)	0.050 [0.001, 0.249] (20)	0.050 [0.001, 0.249] (20)
P2 Med – P1 Low	0.250 [0.087, 0.491] (20)	0.095 [0.012, 0.304] (21)	0.000 [0.000, 0.161] (21)
P2 Med – P1 Med	0.262 [0.139, 0.420] (42)	0.116 [0.039, 0.251] (43)	0.216 [0.098, 0.382] (37)
P2 Med – P1 High	0.333 [0.099, 0.651] (12)	0.200 [0.043, 0.481] (15)	0.238 [0.082, 0.472] (21)
P2 High – P1 Low	0.000 [0.000, 0.195] (17)	0.100 [0.012, 0.317] (20)	0.172 [0.058, 0.358] (29)
P2 High – P1 Med	0.130 [0.028, 0.336] (23)	0.083 [0.010, 0.270] (24)	0.125 [0.016, 0.383] (16)
P2 High – P1 High	0.250 [0.107, 0.449] (28)	0.162 [0.062, 0.320] (37)	0.083 [0.018, 0.225] (36)

P2-P1 pairing	Report 4	Report 5	Report 6
P2 Low – P1 Low	0.167 [0.075, 0.302] (48)	0.250 [0.155, 0.366] (72)	0.029 [0.001, 0.153] (34)
P2 Low – P1 Med	0.250 [0.107, 0.449] (28)	0.371 [0.215, 0.551] (35)	0.154 [0.019, 0.454] (13)
P2 Low – P1 High	0.150 [0.032, 0.379] (20)	0.286 [0.173, 0.422] (56)	0.111 [0.014, 0.347] (18)
P2 Med – P1 Low	0.083 [0.010, 0.270] (24)	0.111 [0.024, 0.292] (27)	0.161 [0.055, 0.337] (31)
P2 Med – P1 Med	0.094 [0.031, 0.207] (53)	0.269 [0.175, 0.382] (78)	0.200 [0.084, 0.369] (35)
P2 Med – P1 High	0.286 [0.113, 0.522] (21)	0.375 [0.249, 0.515] (56)	0.158 [0.034, 0.396] (19)
P2 High – P1 Low	0.077 [0.009, 0.251] (26)	0.194 [0.082, 0.360] (36)	0.312 [0.110, 0.587] (16)
P2 High – P1 Med	0.250 [0.098, 0.467] (24)	0.258 [0.119, 0.446] (31)	0.038 [0.001, 0.196] (26)
P2 High – P1 High	0.206 [0.087, 0.379] (34)	0.254 [0.177, 0.344] (114)	0.256 [0.130, 0.421] (39)

Notes: Each entry is the mean punishment rate with exact (Clopper-Pearson) 95% confidence interval and cell size N. Example: 0.103 [0.029, 0.242] (39) means a 10.3% punishment rate, 95% CI 2.9%–24.2%, from $N = 39$ observations.

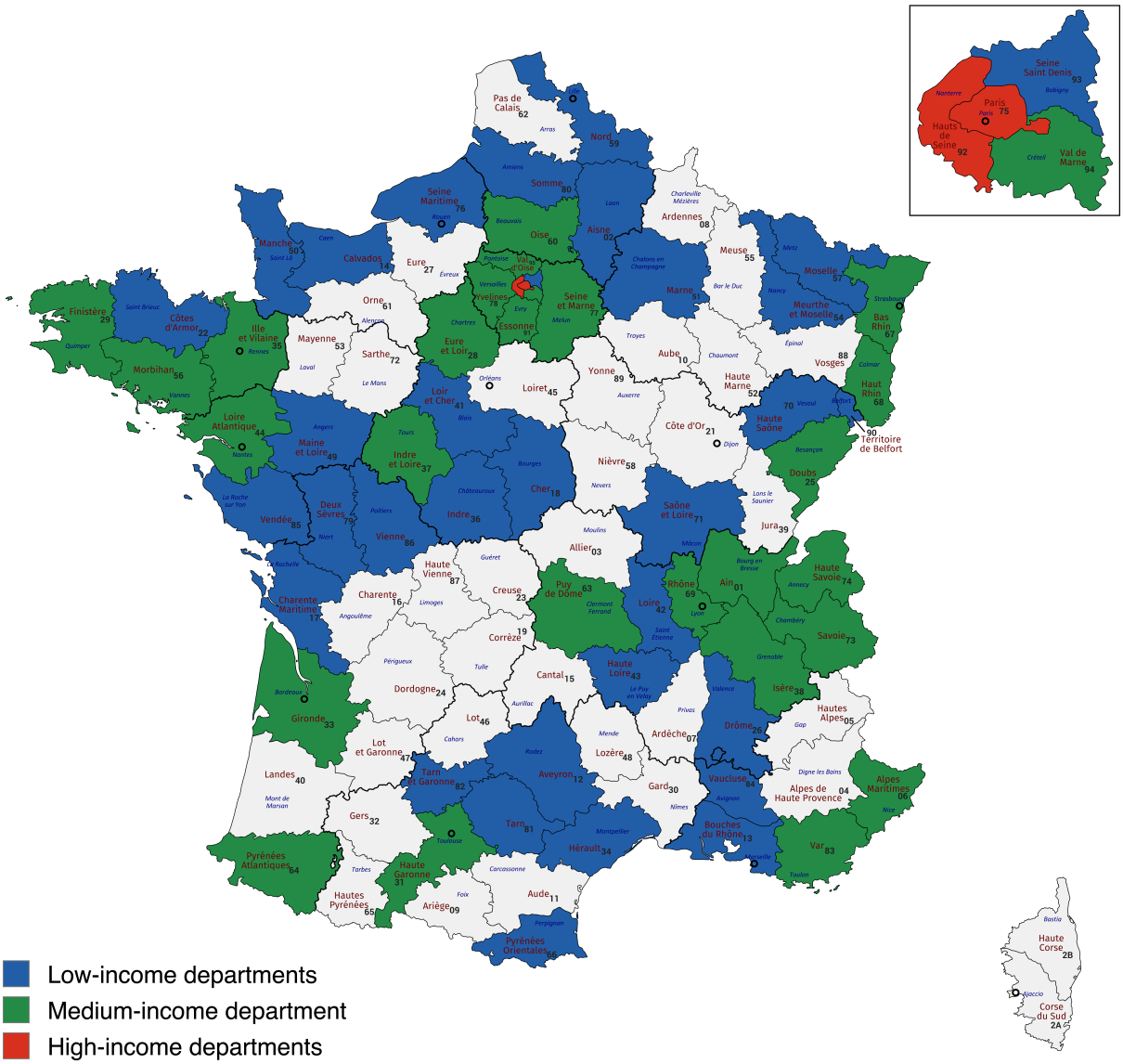
Table 3.33: Reward by P2 SES × P1 SES and reported value: rate [95% CI]

P2–P1 pairing	Report 1	Report 2	Report 3
P2 Low – P1 Low	0.872 [0.726, 0.957] (39)	0.912 [0.763, 0.981] (34)	0.957 [0.852, 0.995] (46)
P2 Low – P1 Med	0.850 [0.621, 0.968] (20)	0.750 [0.476, 0.927] (16)	0.793 [0.603, 0.920] (29)
P2 Low – P1 High	0.950 [0.751, 0.999] (20)	0.762 [0.528, 0.918] (21)	0.767 [0.577, 0.901] (30)
P2 Med – P1 Low	0.909 [0.708, 0.989] (22)	0.920 [0.740, 0.990] (25)	0.636 [0.407, 0.828] (22)
P2 Med – P1 Med	0.828 [0.642, 0.942] (29)	0.837 [0.703, 0.927] (49)	0.757 [0.588, 0.882] (37)
P2 Med – P1 High	0.857 [0.673, 0.960] (28)	0.812 [0.544, 0.960] (16)	0.765 [0.501, 0.932] (17)
P2 High – P1 Low	0.857 [0.572, 0.982] (14)	1.000 [0.858, 1.000] (24)	0.875 [0.676, 0.973] (24)
P2 High – P1 Med	0.778 [0.524, 0.936] (18)	0.714 [0.419, 0.916] (14)	0.588 [0.329, 0.816] (17)
P2 High – P1 High	0.706 [0.525, 0.849] (34)	0.812 [0.636, 0.928] (32)	0.756 [0.597, 0.876] (41)

P2–P1 pairing	Report 4	Report 5	Report 6
P2 Low – P1 Low	0.647 [0.501, 0.776] (51)	0.726 [0.609, 0.824] (73)	0.689 [0.534, 0.818] (45)
P2 Low – P1 Med	0.714 [0.478, 0.887] (21)	0.550 [0.385, 0.707] (40)	0.444 [0.215, 0.692] (18)
P2 Low – P1 High	0.714 [0.537, 0.854] (35)	0.667 [0.447, 0.844] (24)	0.857 [0.572, 0.982] (14)
P2 Med – P1 Low	0.680 [0.465, 0.851] (25)	0.538 [0.334, 0.734] (26)	0.750 [0.533, 0.902] (24)
P2 Med – P1 Med	0.652 [0.498, 0.786] (46)	0.464 [0.355, 0.576] (84)	0.698 [0.539, 0.828] (43)
P2 Med – P1 High	0.684 [0.434, 0.874] (19)	0.654 [0.509, 0.780] (52)	0.667 [0.349, 0.901] (12)
P2 High – P1 Low	0.792 [0.578, 0.929] (24)	0.622 [0.448, 0.775] (37)	0.619 [0.384, 0.819] (21)
P2 High – P1 Med	0.542 [0.328, 0.744] (24)	0.513 [0.348, 0.676] (39)	0.562 [0.377, 0.736] (32)
P2 High – P1 High	0.655 [0.519, 0.775] (58)	0.461 [0.354, 0.570] (89)	0.529 [0.351, 0.702] (34)

Notes: Each entry is the mean reward rate with exact (Clopper-Pearson) 95% confidence interval and cell size N.

Figure 3.27: Participants' departments of residence by SES



The map shows the departments where participants registered for the experiment by SES level:

Low-income: Aisne, Aveyron, Bouches-du-Rhone, Calvados, Charente-Maritime, Cher, Cotes-d'Armor, Deux-Sèvres, Drome, Eure-et-Loir, Haute-Loire, Haute-Saône, Haute-Savoie, Hérault, Indre, Isère, Loir-et-Cher, Loire, Maine-et-Loire, Manche, Marne, Meurthe-et-Moselle, Moselle, Nord, Pas-de-Calais, Pyrénées-Orientales, Saône-et-Loire, Seine-Maritime, Seine-Saint-Denis, Somme, Tarn, Tarn-et-Garonne, Territoire de Belfort, Vaucluse, Vendée, Vienne.

Medium-income: Ain, Alpes-Maritimes, Bas-Rhin, Doubs, Essonne, Finistère, Gironde, Haut-Rhin, Haute-Garonne, Indre-et-Loire, Loire-Atlantique, Morbihan, Oise, Puy-de-Dôme, Pyrenées-Atlantiques, Rhône, Savoie, Seine-et-Marne, Val-d'Oise, Val-de-Marne, Var, Yvelines, Ille-et-Vilaine.

High-income: Hauts-de-Seine, Paris.

Figure 3.28: Originally registered protocol

Favouritism or fairness? Socioeconomic closeness and moral judgments of dishones (#236581)

Author(s)
Irving Argaez (Centre d'Économie de la Sorbonne) - irving.argaez@univ-paris1.fr
Jean-Christophe Vergnaud (Paris 1 Pantheon Sorbonne) - Jean-Christophe.Vergnaud@univ-paris1.fr
Béatrice Boulu-Reshef (Université d'Orléans) - beatrice.boulu-reshef@univ-orleans.fr

Pre-registered on: 2025/07/03 - 02:41 AM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
We investigate whether dishonest behaviour (cheating) is punished more when committed by a socially distant individual, and whether honest behaviour (non-cheating) is more likely to be rewarded when shown by a socially close individual. We test these hypotheses: (H1a) Participants punish more when cheating is committed by a socially distant counterpart; (H1b) Socially distant participants are punished more regardless of their behaviour; (H2a) Participants reward more when non-cheating is committed by a socially close counterpart; and (H2b) Socially close participants are rewarded more regardless of their behaviour.

3) Describe the key dependent variable(s) specifying how they will be measured.
The propensity to cheat in a Die-under-the-cup task (DUTC), where results above the statistical average of 2.5 indicate cheating (die-rolls yield equivalent payoffs in euros, except for 6 that yields a 0€ payoff), and the observer's decision to punish or reward. We compute social closeness using average monthly income in the department of residence (objective measure) and participants' self-reported income (subjective measure). Pairings are determined as close if they share socioeconomic levels and distant otherwise.

4) How many and which conditions will participants be assigned to?
Participants rotate with 12 close and 12 distant counterparts in a between-subject DUTC for 24 rounds. They are assigned to one of two fixed roles: Die-roller (P1) or Observer (P2).There are 3 treatments:
Treatment 1: Punishment
P1 rolls a die, sees the value and reports it; cheating occurs if the values are not identical.
P2 observes the true and reported values and decides whether to punish P1 by deducting -2€ from their payoff. P2 registers their own payoff.
Treatment 2: Reward
P1 same as T1
P2 observes the true and reported values and decides whether to reward P1 by adding +2€ to their payoff. P2 registers payoff.
Treatment 3: Mixed
P1 same as T1.
P2 is assigned to one of three conditions: Observer (no action), Punish (can deduct -2€) or Reward (can add +2€) P2 registers payoff.
P2 are informed about the soecionomic level in the department of residence of P1. P1 knows P2 is assigned to one of the three conditions.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We estimate regressions using the following variables: P1_cheat (binary), P2_cheat, P2_punish (binary), P2_reward (binary), closeness (binary), Treatment (continuous)
H1: Punishment (T1 and T3): regress P2_punish P1_cheat closeness P1_cheat#closeness controls, cluster(participant_id)
H2: Reward (T2 and T3): regress P2_reward P1_cheat closeness P1_cheat#closeness controls, cluster(participant_id)
Treatment Effects (T3): regress P1_cheat Treatment3 closeness Treatment3#closeness controls, cluster(participant_id)
Psychological proxy for P2's own cheating: regress P2_cheat P1_cheat P2_punish P2_reward closeness /// P1_cheat#P2_punish P1_cheat#P2_reward controls, cluster(participant_id)

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
na

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
n = 720 participants: T1 and T2: 144 participants × 24 rounds = 3,456 observations per treatment, for a total of 6,912 observations (3,456 for P2). T3: 432 participants × 24 rounds = 10,368 observations (split by observer, punish and reward = 3,456 each) (5,184 for P2).

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Nothing else to pre-register.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4):1115–1153.

Accinelli, E. and Carrera, E. J. (2012). Corruption driven by imitative behavior. *Economics Letters*, 117(1):84–87.

Acedo-Carmona, C. and Gomila, A. (2014). Personal trust increases cooperation beyond general trust. *PLoS ONE*, 9(8):1–10.

Ahloy, J. and Hamman, J. R. (2019). Personality Traits and Endogenous Group Formation. *Source: Revue économique*, 70(6):999–1020.

Akerlof, G. A. and Kranton, R. E. (1997). Social Distance and Social Decisions. Technical Report 5.

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

Akerlof, G. A. and Kranton, R. E. (2002). Identity and Schooling: Some Lessons for the Economics of Education. *Journal of Economic Literature*, 40:1167–1201.

Alfonso-Costillo, A., Brañas-Garza, P., and López-Martín, M. C. (2022). Does the die-under-the-cup device exaggerate cheating? *Economics Letters*, 214.

Amir, A., Kogut, T., and Bereby-Meyer, Y. (2016). Careful cheating: People cheat groups rather than individuals. *Frontiers in Psychology*, 7(371):1–8.

Anvari, F., Wenzel, M., Woodyatt, L., and Haslam, S. A. (2019). The social psychology of whistleblowing: An integrated model. *Organizational Psychology Review*, 9(1):41–67.

Aquino, K. and Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6):1423–1440.

Aron, A., Aron, E., 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):596–612.

Artinger, F., Exadaktylos, F., Koppel, H., and Sääksvuori, L. (2010). Applying Quadratic Scoring Rule transparently in multiple choice settings: A note. *Working Paper*, (January):1–15.

Ashton, M. C. and Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2):150–166.

Ashton, M. C. and Lee, K. (2008). The prediction of Honesty-Humility-related criteria by the HEXACO and Five-Factor Models of personality. *Journal of Research in Personality*, 42(5):1216–1228.

Baader, M., Starmer, C., Tufano, F., and Gächter, S. (2024). Introducing IOS11 as an extended interactive version of the 'Inclusion of Other in the Self' scale to estimate relationship closeness. *Scientific Reports*, 14(1).

Baccini, E. and Hartmann, S. (2022). The Myside Bias in Argument Evaluation: A

Bayesian Model. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022*, pages 1512–1518.

Bäker, A. . and Mechtel, M. (2015). Peer Settings Induce Cheating on Task Performance.

Balliet, D., Wu, J., and De Dreu, C. K. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychological bulletin*, 140(6):1556–1581.

Barranti, M., Carlson, E. N., and Furr, R. M. (2016). Disagreement About Moral Character Is Linked to Interpersonal Costs. *Social Psychological and Personality Science*, 7(8):806–817.

Bartels, D. M. and Burnett, R. C. (2011). A group construal account of drop-in-the-bucket thinking in policy preference and moral judgment. *Journal of Experimental Social Psychology*, 47(1):50–57.

Beer, A. and Watson, D. (2008). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment*, 90(3):250–260.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.

Benistant, J., Galeotti, F., and Villeval, M. C. (2021). The Distinct Impact of Information and Incentives on Cheating The Distinct Impact of Information and Incentives on Cheating *. Technical report.

Benistant, J., Galeotti, F., and Villeval, M. C. (2022). Competition, information, and the erosion of morals. *Journal of Economic Behavior and Organization*, 204:148–163.

Beranek, B. and Castillo, G. (2022). Continuous Inclusion of Other in the Self. Technical report.

Berg, A. (2019). Identity in economics: a review.

Bernard, M., Hett, F., and Mechtel, M. (2016). Social identity and social free-riding. *European Economic Review*, 90:4–17.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, 132:59–72.

Bilancini, E., Boncinelli, L., Capraro, V., Celadin, T., and Di Paolo, R. (2020). "Do the right thing" for whom? An experiment on ingroup favouritism, group assorting and moral suasion. *Judgment and Decision Making*, 15(2):182–192.

Blanken, I., van de Ven, N., and Zeelenberg, M. (2015). A Meta-Analytic Review of Moral Licensing. *Personality and Social Psychology Bulletin*, 41(4):540–558.

Blokland, T. (2012). Blaming neither the undeserving poor nor the revanchist middle classes: A relational approach to marginalization. *Urban Geography*, 33(4):488–507.

Bone, J. E., McAuliffe, K., and Raihani, N. J. (2016). Exploring the motivations for punishment: Framing and country-level effects. *PLoS ONE*, 11(8).

Boone, C., Declerck, C., and Kiyonari, T. (2010). Inducing Cooperative Behavior among Proselfs versus Prosocials: The Moderating Role of Incentives and Trust. *Journal of Conflict Resolution*, 54(5):799–824.

Bose, N. and Sgroi, D. (2022). The role of personality beliefs and "small talk" in strategic behaviour. *PLoS ONE*, 17(9 September).

Brown-Iannuzzi, J. L., Lundberg, K. B., and McKee, S. E. (2021). Economic inequality and socioeconomic ranking inform attitudes toward redistribution. *Journal of Experimental Social Psychology*, 96.

Bussolo, M., Lebrand, M., and Torre, I. (2020). Feeling Poor, Feeling Rich, or Feeling Middle-Class An Empirical Investigation.

Cameron, L., Chaudhuri, A., Erkal, N., and Gangadharan, L. (2009). Propensities to engage in and punish corrupt behavior: Experimental evidence from Australia, India, Indonesia and Singapore. *Journal of Public Economics*, 93(7-8):843–851.

Castillo, G. (2021). Preference reversals with social distances. *Journal of Economic Psychology*, 86.

Castro Santa, J., Exadaktylos, F., and Soto-Faraco, S. (2018). Beliefs about others' intentions determine whether cooperation is the faster choice. *Scientific Reports*, 8(1):1–10.

Chae, J., Kim, K., Kim, Y., Lim, G., Kim, D., and Kim, H. (2022). Ingroup favoritism overrides fairness when resources are limited. *Scientific Reports*, 12(1).

Charness, G. (2000). Self-serving cheap talk: A test of aumann's conjecture. *Games and Economic Behavior*, 33(2):177–194.

Charroin, L., Fortin, B., Villeval, M. C., Boucher, V., Bramoullé, Y., Chen, Y., Cohn, A., Davidson, R., Fluet, C., Marchand, S., and Shearer, B. (2021). Homophily, Peer Effects, and Dishonesty Homophily, Peer Effects, and Dishonesty *. Technical report.

Chierchia, G. and Coricelli, G. (2015). The impact of perceived similarity on tacit coordination: Propensity for matching and aversion to decoupling choices. *Frontiers in Behavioral Neuroscience*, 9(JULY).

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., and Neuberg, S. L. (1997). Reinterpreting the Empathy-Altruism Relationship: When One Into One Equals Oneness. Technical report.

Conrads, J., Irlenbusch, B., Rilke, R. M., and Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, 34:1–7.

Cooper, J. (2019). Cognitive dissonance: Where we've been and where we're going. *International Review of Social Psychology*, 32(1).

Cooper, W. H. and Withey, M. J. (2009). The strong situation hypothesis. *Personality and Social Psychology Review*, 13(1):62–72.

Costa, P. (1992). Neo PI-R professional manual GWAS of Personality View project Lifespan and Intergenerational Effects of Childhood Malnutrition View project. Technical report.

Costa, P. T. and McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment: Volume 2 -*

Personality Measurement and Testing, pages 179–198. SAGE Publications Inc.

Coyne, I. and Bartram, D. (2002). Assessing the Effectiveness of Integrity Tests: A Review. *International Journal of Testing*, 2(1):15–34.

Crede, A.-K. and von Bieberstein, F. (2020). Reputation and lying aversion in the die roll paradigm: Reducing ambiguity fosters honest behavior. *Managerial and Decision Economics*, 41(4).

Crowe, M. L., Lynam, D. R., and Miller, J. D. (2018). Uncovering the structure of agreeableness from self-report measures. *Journal of Personality*, 86(5):771–787.

Currarini, S. and Mengel, F. (2016). Identity, homophily and in-group bias. *European Economic Review*, 90:40–55.

Dalton, R. (2016). Party identification and its implications. *Oxford Research Encyclopedia of Politics*.

de Dreu, C. K. (2010). Social value orientation moderates ingroup love but not outgroup hate in competitive intergroup conflict. *Group Processes & Intergroup Relations*, 13(6):701–713.

De Freitas, J., Thomas, K., DeScioli, P., and Pinker, S. (2019). Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences of the United States of America*, 116(28):13751–13758.

Deutchman, P., Bračič, M., Raihani, N., and McAuliffe, K. (2021). Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evolution and Human Behavior*, 42(1):12–20.

Devetag, G. and Ortmann, A. (2007). When and why? A critical survey on coordination failure in the laboratory. *Experimental Economics*, 10(3):331–344.

Dimant, E. (2019). Contagion of pro- and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73:66–88.

Drouvelis, M. and Georgantzis, N. (2019). Does revealing personality data affect prosocial behaviour? *Journal of Economic Behavior and Organization*, 159:409–420.

Du, H., Chen, A., Chi, P., and King, R. B. (2020). Preprint of "Income Inequality Reduces Civic Honesty".

Dungan, J. A., Young, L., and Waytz, A. (2019). The power of moral concerns in predicting whistleblowing decisions. *Journal of Experimental Social Psychology*, 85.

Easterbrook, M. J., Hadden, I. R., and Nieuwenhuis, M. (2019). Identities in context: How social class shapes inequalities in education. In *The Social Psychology of Inequality*, pages 103–121. Springer International Publishing.

Easterbrook, M. J., Kuppens, T., and Manstead, A. S. (2020). Socioeconomic status and the structure of the self-concept. *British Journal of Social Psychology*, 59(1):66–86.

Elbæk, C. T., Mitkidis, P., Aarøe, L., and Otterbring, T. (2023). Subjective socioeconomic status and income inequality are associated with self-reported morality across 67 countries. *Nature Communications*, 14(1).

Ellemers, N., Pagliaro, S., Barreto, M., and Leach, C. W. (2008). Is It Better to Be Moral

Than Smart? The Effects of Morality and Competence Norms on the Decision to Work at Group Status Improvement. *Journal of Personality and Social Psychology*, 95(6):1397–1410.

Fagbenro, D. A. (2019). Personality Traits and Attitude toward Corruption among Government Workers. *Psychology and Behavioral Science International Journal*, 11(1).

Falk, A. and Zimmermann, F. (2024). Attention and Dread: Experimental Evidence on Preferences for Information. *Management Science*, 70(10):7090–7100.

Fan, C. S., Wei, X., Wu, J., and Zhang, J. (2022). Observability and peer effects: Theory and evidence from a field experiment. *Journal of Economic Behavior and Organization*, 200:847–867.

Fehr, D., Kübler, D., and Danz, D. (2008). Information and Beliefs in a Repeated Normal-form game. *Philosophy of Information*, (3627):551–577.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise-an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Fries, T., Gneezy, U., Kajackaite, A., and Parra, D. (2021). Observability and lying. *Journal of Economic Behavior and Organization*, 189:132–149.

Gächter, S. and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595):496–499.

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).

Galeotti, F., Rilke, R. M., and Verrina, E. (2024). Beliefs and Group Dishonesty: The Role of Strategic Interaction and Responsibility. Technical report.

Gibson, R., Tanner, C., and Wagner Alexander F. (2013). Preferences for Truthfulness: }Heterogeneity Among and Within Individuals. *American Economic Review*, 103(1):532–548.

Gino, F., Ayal, S., and Ariely, D. (2009). Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel. Technical Report 3.

Gino, F., Ayal, S., and Ariely, D. (2012). Self-Serving Altruism? When Unethical Actions That Benefit Others Do Not Trigger Guilt. Technical report.

Giorgetta, C., Grecucci, A., Graffeo, M., Bonini, N., Ferrario, R., and Sanfey, A. G. (2021). Expect the Worst ! Expectations and Social Interactive Decision Making. *Brain Sciences*, 11(572).

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Goldstein, N. J. and Cialdini, R. B. (2007). The spyglass self: A model of vicarious self-perception. *Journal of Personality and Social Psychology*, 92(3):402–417.

Greenberg, S. and Org, C. (2018). Calibration Scoring Rules for Practical Prediction Training. Technical report.

Gries, T., Müller, V., and Jost, J. T. (2022). The Market for Belief Systems: A Formal

Model of Ideological Choice. *Psychological Inquiry*, 33(2):65–83.

Grigoryan, L. (2020). Crossed categorization outside the lab: Findings from a factorial survey experiment. *European Journal of Social Psychology*, 50(5):983–1000.

Grigoryan, L., Seo, S., Simunovic, D., and Hofmann, W. (2023). Helping the ingroup versus harming the outgroup: Evidence from morality-based groups. *Journal of Experimental Social Psychology*, 105.

Gross, J. and De Dreu, C. K. (2021). Rule Following Mitigates Collaborative Cheating and Facilitates the Spreading of Honesty Within Groups. *Personality and Social Psychology Bulletin*, 47(3):395–409.

Gross, J., Leib, M., Offerman, T., and Shalvi, S. (2018). Ethical Free Riding: When Honest People Find Dishonest Partners. *Psychological Science*, 29(12):1956–1968.

Gueguen, N., Jacob, C., and Martin, A. (2009). Mimicry in social interaction: Its effect on human judgment and behavior. *European Journal of Social Sciences*, 8(2):253–259.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M., Lagos, P., Norris, E., Ponarin, and B. Puranen et al. (eds.) (2020). World Values Survey: Round Seven – Country-Pooled Datafile. Technical report, JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Hauge, L. (2007). Identity and Place: A Critical Comparison of Three Identity Theories.

Hauser, O. P., Kraft-Todd, G. T., Rand, D. G., Nowak, M. A., and Norton, M. I. (2021). Invisible inequality leads to punishing the poor and rewarding the rich. *Behavioural Public Policy*, 5(3):333–353.

Hermann, D. and Ostermaier, A. (2018). Be close to me and I will be honest How social distance influences honesty.

Hershcovis, M. S., Neville, L., Reich, T. C., Christie, A. M., Cortina, L. M., and Shan, J. V. (2017). Witnessing wrongdoing: The effects of observer power on incivility intervention in the workplace. *Organizational Behavior and Human Decision Processes*, 142:45–57.

Hewston, M., Rubin, M., and Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, (53):575–604.

Heyman, T., Vankrunkelsven, H., Voorspoels, W., White, A., Storms, G., and Verheyen, S. (2020). When Cheating is an Honest Mistake: A Critical Evaluation of the Matrix Task as a Measure of Dishonesty. *Collabra: Psychology*, 6(1).

Hilbig, B. E., Hessler, C. M., Thielmann, I., Wührl, J., and Zettler, I. (2015). What lies beneath: How the distance between truth and lie drives dishonesty. *Personality and Individual Differences*, 80(2):263–266.

Hilbig, B. E., Zettler, I., Leist, F., and Heydasch, T. (2013). It takes two: Honesty-Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, 54(5):598–603.

Hoffmann, T. (2013). The Effect of Belief Elicitation on Game Play. pages 1–26.

Holt, C. A. and Laury, S. K. (2002). Risk Aversion and Incentive Effects. Technical

Report 5.

Hughes, B. T., Flournoy, J. C., and Srivastava, S. (2020). Is Perceived Similarity More Than Assumed Similarity?: An Interpersonal Path to Seeing Similarity Between Self and Others. *Journal of Personality and Social Psychology*, 121(1):184–200.

Hyndman, K., Terracol, A., and Vaksmann, J. (2013). Beliefs and (In)Stability in Normal-Form Games. (47221).

Inglehart, R. (2000). Culture and Democracy. In *Culture Matters: How Values Shape Human Progress*, pages 80–97. New York: Basic Books.

INSEE Références (2021). La France et ses territoires. Technical report.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2025). The Origins and Consequences of Affective Polarization in the United States. 53:29.

Jansson, F. (2015). What games support the evolution of an ingroup bias? *Journal of Theoretical Biology*, 373:100–110.

Jansson, F. and Eriksson, K. (2015). Cooperation and shared beliefs about trust in the assurance game. *PLoS ONE*, 10(12):1–13.

John, O., Naumann, L., and Soto, C. (2008). *Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues.* Guilford Press, 3rd edition.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90(12).

Jordan, P. J., Troth, A. C., and Yan, H. (2024). Objective and subjective measurement in applied business settings: Improving research in organizations. *Australian Journal of Management*.

Kajonius, P. J. and Dåderman, A. M. (2014). Exploring the relationship between honesty-humility, the big five, and liberal values in Swedish students. *Europe's Journal of Psychology*, 10(1):104–117.

Kaluza, B., Institute, J. S., Kaminka, G., Tambe, M., Kaluža, B., and Kaminka, G. A. (2012). Detection of suspicious behavior from a sparse set of multiagent interactions. Technical report.

Kang, P., Burke, C. J., Tobler, P. N., and Hein, G. (2021). Why we learn less from observing outgroups. *Journal of Neuroscience*, 41(1):144–152.

Kaushik, M., Singh, V., and Chakravarty, S. (2021). Rewards, Detection and Dishonesty: Experimental Evidence from India. *SSRN Electronic Journal*.

Kim, J. E. and Tsvetkova, M. (2021). Cheating in online gaming spreads through observation and victimization. *Network Science*, 9(4):425–442.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms Make Preferences Social. *Journal of the European Economic Association*, 14(3):608–638.

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2019). Intuitive Honesty Versus Dishonesty: Meta-Analytic Evidence. *Perspectives on Psychological Science*, 14(5):778–796.

Kocher, M., Martinsson, P., and Visser, M. (2012). Social background, cooperative behavior, and norm enforcement. *Journal of Economic Behavior and Organization*, 81(2):341–354.

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, 64(9):3995–4008.

Korbel, V. (2016). Do we lie in groups? An experimental evidence. *Applied Economic Letters*, 24(15):1107–1111.

Kraus, M. W., Piff, P. K., and Keltner, D. (2011). Social class as culture: The convergence of resources and rank in the social realm. *Current Directions in Psychological Science*, 20(4):246–250.

Kreps, D. M. (1992). *Game Theory and Economic Modelling*. Oxford.

Kroher, M. and Wolbring, T. (2015). Social control, social learning, and cheating: Evidence from lab and online experiments on dishonesty. *Social Science Research*, 53:311–324.

Ladley, D., Wilkinson, I., and Young, L. (2015). The impact of individual versus group rewards on work group performance and cooperation: A computational social science approach. *Journal of Business Research*, 68(11):2412–2425.

Lane, T. (2023). The strategic use of social identity CeDEx Discussion Paper Series. Technical report.

Larrouy, L. and Lecouteux, G. (2017). Mindreading and endogenous beliefs in games. *Journal of Economic Methodology*, 24(3):318–343.

Le Coq, C., Tremewan, J., and Wagner, A. K. (2015). On the effects of group identity in strategic environments. *European Economic Review*, 76:239–252.

Lee, J. J., Hardin, A. E., Parmar, B., and Gino, F. (2019). The interpersonal costs of dishonesty: How dishonest behavior reduces individuals' ability to read others' emotions. *Journal of Experimental Psychology: General*, 148(9):1557–1574.

Leib, M., Köbis, N., Soraperra, I., Weisel, O., and Shalvi, S. (2021). Collaborative Dishonesty: A Meta-Analytic Review. *Psychological Bulletin*, 147(12):1241–1268.

Leibbrandt, A., López-Pérez, R., and Spiegelman, E. (2023). Reciprocal, but inequality averse as well? Mixed motives for punishment and reward. *Journal of Economic Behavior and Organization*, 210:91–116.

Leidner, B., Castano, E., Zaiser, E., and Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin*, 36(8):1115–1129.

Lönnqvist, J. E., Ilmarinen, V. J., and Verkasalo, M. (2021). Who likes whom? The interaction between perceiver personality and target look. *Journal of Research in Personality*, 90.

Loustau, T., Glassman, J., Martin, J. W., Young, L., and McAuliffe, K. (2024). The impact of group membership on punishment versus partner rejection. *Scientific Reports*, 14(1).

Lutz, G. and Lauener, L. (2020). Measuring party affiliation. Technical report, Lausanne: Swiss Centre of Expertise in the Social Sciences (FORS)., Lausanne.

Macků, K., Caha, J., Pászto, V., and Tuček, P. (2020). Subjective or objective? How objective measures relate to subjective life satisfaction in Europe. *ISPRS International Journal of Geo-Information*, 9(5).

Magni, G. (2021). Economic inequality, immigrants and selective solidarity: From perceived lack of opportunity to in-group favoritism.

Mann, H., Garcia-Rada, X., Houser, D., and Ariely, D. (2014). Everybody else is doing it: Exploring social transmission of lying behavior. *PLoS ONE*, 9(10).

Manstead, A. S. (2018). The psychology of social class: How socioeconomic status impacts thought, feelings, and behaviour. *British Journal of Social Psychology*, 57(2):267–291.

Martin, R. A. (2015). *Perceived and Actual Similarity as Predictors of Self-Disclosure and Perceived Understanding at Zero Acquaintance*. PhD thesis.

Martinangeli, A. F. and Martinsson, P. (2020). We, the rich: Inequality, identity and cooperation. *Journal of Economic Behavior and Organization*, 178:249–266.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people. *Journal of Marketing Research*, 45(6):633–644.

McFerran, B., Aquino, K., and Duffy, M. (2010). How Personality and Moral Identity Relate to Individuals' Ethical Ideology. *Business Ethics Quarterly*, 20(1):35–56.

Mendoza, S. A., Lane, S. P., and Amodio, D. M. (2014). For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game. *Social Psychological and Personality Science*, 5(6):662–670.

Meyners, J., Barrot, C., Becker, J. U., and Goldenberg, J. (2017). The role of mere closeness: How Geographic proximity affects social influence. *Journal of Marketing*, 81(5):49–66.

Michaeli, M. (2020). Grouping, in-group bias and the cost of cheating. *Games and Economic Behavior*, 121:90–107.

Molho, C., De Petrillo, F., Garfield, Z. H., and Slewe, S. (2024). Cross-societal variation in norm enforcement systems.

Moss, R. H., Kelly, B., Bird, P. K., and Pickett, K. E. (2023). Examining individual social status using the MacArthur Scale of Subjective Social Status: Findings from the Born in Bradford study. *SSM - Population Health*, 23.

OECD (2024). OECD Survey on Drivers of Trust in Public Institutions – 2024 Results: Building Trust in a Complex Policy
}Environment. Technical report, OECD Publishing, Paris.

Offerman, T., Sonnemans, J., van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies*, 76(4):1461–1489.

Owuamalam, C. K., Rubin, M., Spears, R., and Weerabangsa, M. M. a. (2017). Why Do

People from Low-Status Groups Support Class Systems that Disadvantage Them? A Test of Two Mainstream Explanations in Malaysia and Australia. *Journal of Social Issues*, 73(1):80–98.

Panagopoulos, C., Leighley, J. E., and Hamel, B. T. (2017). Are Voters Mobilized by a 'Friend-and-Neighbor'on the Ballot? Evidence from a Field Experiment. *Political Behavior*, 39(4):865–882.

Pansini, R., Campennì, M., and Shi, L. (2018). Asymmetric use of punishment in socioeconomic segregated societies leads to an unequal distribution of wealth. Technical report.

Proto, E., Rustichini, A., Deyoung, C., Friebel, G., Grimalda, G., Isoni, A., Loomes, G., Manzini, P., Mariotti, M., Miller, J., Oswald, A., and Stewart, N. (2014). Cooperation and Personality. Technical report.

Pulfrey, C., Durussel, K., and Butera, F. (2018). The good cheat: Benevolence and the justification of collective cheating. *Journal of Educational Psychology*, 110(6):764–784.

Rantakari, H. (2023). How to reward honesty? *Journal of Economic Behavior and Organization*, 207:129–145.

Régner, I. and Monteil, J.-M. (2007). Low-and high-socioeconomic status students preference for ingroup comparisons and their underpinning ability expectations. *Revue Internationale de Psychologie Sociale*, 20(1):87–104.

Renger, D., Lohmann, J. F., Renger, S., and Martiny, S. E. (2024). Socioeconomic status and self-regard income predicts self-respect over time. *Social Psychology*, 55(1):12–24.

Rijnks, R. H. and Strijker, D. (2013). Spatial effects on the image and identity of a rural area. *Journal of Environmental Psychology*, 36:103–111.

Robalo, P., Schram, A., and Sonnemans, J. (2017). Other-regarding preferences, in-group bias and political participation: An experiment. *Journal of Economic Psychology*, 62:130–154.

Rothstein, B. (2011). Anti-corruption: The indirect 'big bang' approach. *Review of International Political Economy*, 18(2):228–250.

Rothstein, B. and Eek, D. (2009). Political Corruption and Social Trust. *Rationality and Society*, 21(1):81–112.

Rubin, M., Badea, C., and Jetten, J. (2014). Low status groups show in-group favoritism to compensate for their low status and compete for higher status. *Group Processes & Intergroup Relations*, 17(5):563–576.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under "almost common knowledge". *American Economic Review*, 79(3):385–391.

Rubinstein, A. and Salant, Y. (2016). Isn't everyone like me?": On the presence of self-similarity in strategic interactions. *Judgment and Decision Making*, 11(2):168–173.

Ruch, W., Bruntsch, R., and Wagner, L. (2017). The role of character traits in economic

games. *Personality and Individual Differences*, 108:186–190.

Rullo, M., Monaco, S., Giannini, F., Livi, S., and Presaghi, F. (2019). In the name of truth: People's reactions to ingroup and outgroup members who self-disclose a severe error. *Social Science Journal*, 56(3):421–424.

Rullo, M., Presaghi, F., Baldner, C., Livi, S., and Butera, F. (2024). Omertà in intragroup cheating: The role of ingroup identity in dishonesty and whistleblowing. *Group Processes and Intergroup Relations*, 27(1):41–61.

Rustichini, A. (2009). Neuroeconomics: what have we found, and what should we search for.

Rustichini, A., DeYoung, C. G., Anderson, J. E., and Burks, S. V. (2016). Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation. *Journal of Behavioral and Experimental Economics*, 64:122–137.

Ruzzier, C. A. and Woo, M. D. (2023). Discrimination with inaccurate beliefs and confirmation bias. *Journal of Economic Behavior and Organization*, 210:379–390.

Ryvkin, D., Serra, D., and Tremewan, J. (2017). I paid a bribe: An experiment on information sharing and extortionary corruption. *European Economic Review*, 94:1–22.

Schiller, B., Baumgartner, T., and Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior*, 35(3):169–175.

Schram, A., Zheng, J. D., and Zhuravleva, T. (2022). Corruption: A cross-country comparison of contagion and conformism. *Journal of Economic Behavior and Organization*, 193:497–518.

Sgroi, D., Yeo, J., and Zhuo, S. (2021). Ingroup Bias with Multiple Identities: The Case of Religion and Attitudes Towards Government Size. Technical report.

Shalvi, S., Dana, J., Handgraaf, M. J., and De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190.

Siniver, E., Tobol, Y., and Yaniv, G. (2022). Collective Punishment and Cheating in the Die-Under-the-Cup Task. *Experimental Psychology*, 69(1):40–45.

Skyrms, B. (2003). *The Stag Hunt and the Evolution of Social Structure*. Number 1. Cambridge University Press, Cambridge.

Sosa, M. and Maoret, M. (2023). Close to Me: The Impact of the Interplay of Physical and Social Proximity on Dyadic Collaboration Effectiveness. Technical report.

Stahl, D. and Huyck, J. V. (2002). Learning conditional behavior in similar stag hunt games. (January).

Steinel, W., Valtcheva, K., Gross, J., Celse, J., Max, S., and Shalvi, S. (2022). (Dis)honesty in the face of uncertain gains or losses. *Journal of Economic Psychology*, 90.

Tajfel, H. and Turner, J. C. (1979). An integrative theory of intergroup conflict. In *The social psychology of intergroup relations*, pages 33–37. Brooks/Cole, Monterey, CA.

Thielmann, I., Akrami, N., Babarović, T., Belloch, A., Bergh, R., Chirumbolo, A., Čolović, P., de Vries, R. E., Dostál, D., Egorova, M., Gnisci, A., Heydasch, T., Hilbig, B. E., Hsu, K. Y., Izdebski, P., Leone, L., Marcus, B., Međedović, J., Nagy, J., Parshikova, O., Perugini, M., Petrović, B., Romero, E., Sergi, I., Shin, K. H., Smederevac, S., Šverko, I., Szarota, P., Szirmák, Z., Tatar, A., Wakabayashi, A., Wasti, S. A., Záškodná, T., Zettler, I., Ashton, M. C., and Lee, K. (2020). The HEXACO–100 Across 16 Languages: A Large-Scale Test of Measurement Invariance. *Journal of Personality Assessment*, 102(5):714–726.

Thielmann, I., Hilbig, B. E., Klein, S. A., Seidl, A., and Heck, D. W. (2024). Cheating to benefit others? On the relation between Honesty-Humility and prosocial lies. *Journal of Personality*, 92(3):870–882.

Thomas, G. O., Poortinga, W., and Sautkina, E. (2016). The Welsh Single-Use Carrier Bag Charge and behavioural spillover. *Journal of Environmental Psychology*, 47(2880):126–135.

Thomas, K. A., DeScioli, P., Haque, O. S., and Pinker, S. (2014). The Psychology of Coordination and Common Knowledge. *Journal of Personality and Social Psychology*, 107(4):657–676.

Tobol, Y., Siniver, E., and Yaniv, G. (2020). Do tightwads cheat more? Evidence from three field experiments. *Journal of Economic Behavior and Organization*, 180:148–158.

Tsvetkova, M. and Macy, M. W. (2015). The social contagion of antisocial behavior. *Sociological Science*, 2:36–49.

Van Assche, J., Politi, E., Van Dessel, P., and Phalet, K. (2020). To punish or to assist? Divergent reactions to ingroup and outgroup members disobeying social distancing. *British Journal of Social Psychology*, 59(3):594–606.

van de Ven, J. and Villeval, M. C. (2015). Dishonesty under scrutiny. *Journal of the Economic Science Association*, 1(1):86–99.

Van De Walle, S. (2008). *Perceptions of corruption as distrust? Cause and effect in attitudes toward government*. Number June.

Van Huyck, J., Viriyavipart, A., and Brown, A. L. (2018). When less information is good enough: experiments with global stag hunt games. *Experimental Economics*, 21(3):527–548.

Van Huyck, J. B., Battalio, R. C., and Beil, R. O. (1990). Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure. *American Economic Review*, 80(1):234–248.

van Oosten, S. (2025). The Importance of In-group Favoritism in Explaining Voting for PRRPs: A Study of Minority and Majority Groups in France, Germany and the Netherlands. Technical report, European Center for Populism Studies, Brussels.

Volk, S., Thöni, C., and Ruigrok, W. (2011). Personality, personal values and cooperation preferences in public goods games: A longitudinal study. *Personality and Individual Differences*, 50(6):810–815.

Waytz, A., Dungan, J., and Young, L. (2013). The whistleblower's dilemma and the fairness-loyalty tradeoff. *Journal of Experimental Social Psychology*, 49(6):1027–1033.

Weiner, D. S. and Laurent, S. M. (2021). The (Income-Adjusted) Price of Good Behavior: Documenting the Counter-Intuitive, Wealth-Based Moral Judgment Gap. *Journal of Experimental Psychology: General*, 150(3):484–506.

Weisel, O. and Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences of the United States of America*, 112(34):10651–10656.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, 115(2):157–168.

Windrich, I., Kierspel, S., Neumann, T., Berger, R., and Vogt, B. (2024). Enforcement of Fairness Norms by Punishment: A Comparison of Gains and Losses. *Behavioral Sciences*, 14(1).

Winter, F. and Zhang, N. (2018). Social norm enforcement in ethnically diverse communities. 115(11):2722–2727.

Wu, J., Balliet, D., and Van Lange, P. A. (2016). Reputation, Gossip, and Human Cooperation.

Zhao, K. and Smillie, L. D. (2015). The Role of Interpersonal Traits in Social Decision Making: Exploring Sources of Behavioral Heterogeneity in Economic Games. *Personality and Social Psychology Review*, 19(3):277–302.

Zhou, L., Su, C., Sun, X., Zhao, X., and Choo, K. K. R. (2018). Stag hunt and trust emergence in social networks. *Future Generation Computer Systems*, 88:168–172.

General conclusions

The study of social identity has long bridged psychological and economic approaches. Research in both disciplines highlights the multiple psychological, social and economic benefits that individuals obtain from belonging to social groups. Seminal contributions by Akerlof and Kranton show that identity shapes preferences and behaviour: an individual's payoffs depend not only on material outcomes, but also on their individual's preferences and the different social categorisations they inhabit Akerlof and Kranton (1997, 2000) Akerlof and Kranton (2002).

Since then, the literature has proliferated across several subfields of economics—from organisational and behavioural economics to more niche branches venturing into the economics of discrimination (Berg, 2019) or neuroeconomics (Rustichini, 2009). A central insight in this recent work is that identity salience is often endogenous: agents, as well as third parties, can strategically activate particular identity dimensions. When identity cues are malleable and the payoff consequences of identification are salient, individuals may deliberately select the identity dimension along which an interaction takes place, thereby shaping strategic behaviour and economic outcomes (Lane, 2023).

Against this backdrop, this thesis investigated whether disclosing social information about others and about oneself re-orients preferences and behaviour in strategic interactions. We show that making social information visible has ambivalent effects: visibility can reduce uncertainty, improve predictability, and support coordination, but it can also distort beliefs and induce selective treatment.

More specifically, across three experimental chapters, we find that (1) trait disclosure shifts beliefs and tends to push play toward a safer, less cooperative equilibrium in trust-sensitive settings; (2) measures of social closeness, namely socioeconomic status and political alignment, systematically alter expectations but do not robustly increase dishonesty in an anonymous DUTC; and (3) third-party punishment and reward are socially selective, with observers showing greater leniency toward socially close individuals and, in the case of punishment, socioeconomic differences moderating sanction severity.

We deem these analyses relevant for different reasons. First, while the concept of identity economics (i.e., who we are has an impact on our economic decisions) is not new, it has reclaimed a prominent space in the research in light of the increasing political and socioeconomic divisions facing societies. This is therefore a unifying concept across the three chapters: predicting behaviour by incorporating one's sense of self into decision-making.

Second, the thesis speaks directly to the mechanisms through which social information shapes behaviour. A growing empirical and theoretical literature shows that disclosure, visibility and identity cues go beyond assigning labels: they change beliefs, expectations and the structure of incentives, often in context-dependent ways. From a policy and organisational perspective, this implies that decisions about transparency and disclosure are not neutral, as revealing social information can improve coordination where goals are aligned but can also institutionalise selective treatment and unequal sanctioning when identities map onto status differences.

In what follows, we present a brief summary of the main findings in the thesis, how they map our research questions, the limitations we found in our study and its contributions. We finalise with the policy implications from the thesis and where its potential lies for future research.

Chapter-by-Chapter Findings

Chapter 1 shows that making personality labels visible increases predictability but can undermine cooperation. When revealing whether participants were trusting or mistrusting personality types, first-order beliefs shift toward trusting types and overall predictability rises; however, their choices drift toward the safer, less efficient equilibrium in the stag-hunt. Overall, the chapter showed that visibility increases alignment of expectations but does not reliably improve cooperative efficiency.

In the experiment, cooperation fell sharply and progressively as label visibility increased, indicating that the more social information was revealed, the more unlikely cooperation seemed. Moreover, this effect appeared as a broad treatment effect rather than one concentrated either personality type. While descriptive statistics hinted at some type-specific patterns, multivariate analysis showed that the main driver was a change in strategic incentives - players converged on a safe equilibrium because revised beliefs raised the perceived downside of cooperative choices, not because they adopted the disclosed label as a new social identity or because they systematically discriminated others who did not share this label.

Moreover, the chapter set out to test four specific mechanisms that conceptualised these effects. Results showed that (a) Self-identification (Mechanism 1) had no detectable effect on cooperative choices. Participants did not internalise the label in a way that could change their intrinsic willingness to cooperate; (b) Type-based discrimination and homophily (Mechanism 2) were not significant in regression analyses, as observed descriptive patterns were not driven by persistent preference-based favouritism or exclusion; (c) First-order beliefs (Mechanism 3) were the strongest and most consistent predictor of cooperation. Labels primarily operated by changing what players expect of their partners; and (d) Second-order beliefs (Mechanism 4) had limited effects and, where present, acted mainly through their impact on counterparts' first-order beliefs.

The key takeaway from this chapter is that beliefs dominated preferences and that label visibility operated by distorting expectations rather than by shifting intrinsic preferences or generating durable type-based discrimination. In this sense, labels changed what people expected more than they changed their intrinsic preferences to cooperate with others.

Chapter 2 tested whether social proximity, operationalised via socioeconomic status and political alignment, influences misreporting payoffs in a DUTC and whether observation by socially close counterparts amplifies this behaviour. Our analyses consistently found little evidence that closeness systematically increases dishonest reporting. Moreover, the chapter introduced an objective and a subjective measure of closeness, with the former showing, at best, weak and context-dependent associations, while the latter is consistently null.

Results from this chapter suggest that everyday forms of social proximity do not directly entice cheating behaviour, nor in-group cheating incentives when payoffs are neutral. While effects were weaker than anticipated, with data indicating that simply pairing similar individuals does not systematically increase dishonesty under passive monitoring, the chapter did shed light on certain aspects that we deem worth noting.

First, the study offered a multi-scale operationalisation of social closeness and a clean obser-

vation design at the individual level, enabling a clear separation of between-subject differences from within-person behavioural changes. Second, it presented two real-world dimensions to model social closeness that, while not significantly explicative of cheating behaviour, were effective in conceptualising participants' perceptions of closeness. While the measurement of both, socioeconomic status and political affiliation, can certainly be improved in future studies, their use can certainly be replicated.

Chapter 3 examined whether social closeness and socioeconomic status shape punishment and reward decisions in the DUTC. Our results showed that norm enforcement is predominantly evidence-driven: punishment increased with suspicion, while rewards increased as suspicion declined. Social closeness mediated these effects, with objectively close pairs being punished less and rewarded more at comparable suspicion levels. Importantly, low-SES individuals were punished more harshly, while high-SES participants faced stricter standards among similarly high-status observers.

Similarly, subjective closeness affected responsiveness, softening the decline in rewards with increasing suspicion but not systematically shifting baselines. This demonstrates that selective norm enforcement is simultaneously evidence-based and socially filtered: proximity biases the conversion of identical signals into sanctions or rewards, while status cues further modulate expectations and responses.

Contributions and Originality

The thesis makes some notable contributions:

- We present a unified framework that analyses the impact of social information disclosure from beliefs, to individual behaviour, to third-party responses, filling an existing gap in the literature as these are usually studied in isolation.
- We provide an empirical tests of belief-based distortions triggered by visible traits, demonstrating first-order belief bias and second-order pessimism in a strategic game.
- We introduce multi-scale measures of objective and subjective closeness, allowing a nuanced understanding of in-group effects, capturing both baseline shifts and responsiveness differences in norm enforcement.
- In Chapter 3, we introduce a suspicion-based index linking observed behaviour to observer actions, enabling clean comparisons between punishment and reward decisions in a controlled yet realistic setting.
- Across chapters, we show that expectations mediate behaviour, whether in coordination, dishonesty, or norm enforcement, highlighting the centrality of epistemic processes in strategic environments.
- By combining repeated interaction designs with individual-level observation measures, the thesis enables a richer understanding of micro-mechanisms governing cooperation, cheating, and sanctions.

Limitations and Future Directions

We identify some caveats and limitations in how the chapters were executed. First, we acknowledge different constraints in our experimental designs: the misclassification of types on the hybrid personality inventory or in the economic and political classifications could have reduced the effects we anticipated in these variables. Future research could integrating richer identity, status, and relationship measures to capture subtler forms of social influence.

Also, key aspects of how out settings were implemented, such as risk perceptions arising from label visibility, neutral-payoffs or the nature of the economic games selected may have attenuated true behavioural effects. We think in particular of the payoff structure of the stag-hunt game shifting preferences towards safe choices, thereby reducing the effects of our independent and dependent variables. Similarly, our implementation of the Die-under-the-cup task could have showed a contagion effect in cheating that tainted observers decisions. Future work should learn from these underlying effects to counter diminished returns in the significance of results.

Furthermore, the external validity of chapters 2 and 3 in particular, might be limited to specific SES and political framings in online DUTCs. Field or organisational contexts with salient group objectives, dependent payoffs, or norm enforcement mechanisms could reveal stronger effects.

Policy implications

The main implication of this thesis is that transparency and visibility are not unambiguously beneficial. Revealing identity or personality labels tends to increase predictability, however, can also undermine cooperative efficiency. Where institutions treat disclosure as an automatic good, they risk replacing mutually beneficial coordination with safer, less efficient behaviours. Consequently, policy design should treat transparency as an instrument that must be carefully bounded and balanced with mechanisms that privilege observable behaviour over static labels.

The findings also caution against relying on everyday social proximity as a means of shaping honesty or compliance. In our experiments, passive social closeness did not reliably increase or decrease dishonest behaviour; instead, deterrence was driven primarily by the clarity of monitoring. This suggests that randomised checks or awareness of sanctioning protocols are likely to be more effective at reducing dishonest behaviour than attempts to manipulate social cues or proximity.

Our results also shed light on the importance of reward and recognition schemes to avoid reinforcing identity-based advantages. Public acknowledgement and incentives are most fair and effective when they are tied to concrete, verifiable actions, rather than to perceived traits or affiliation. Where possible, automating rewards according to transparent rules reduces discretionary bias and the scope for identity-driven favouritism.