

IMT School for Advanced Studies, Lucca
Lucca, Italy

PhD thesis co-supervised by the Universidad de Alicante
Alicante, Spain

Three Essays on Behavioral Game Theory

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

By

Juan Francisco Blázquez Pulido

2026

The dissertation of Juan Francisco Blázquez Pulido is approved.

PhD Program Coordinator: Ennio Bilancini, IMT School for Advanced Studies Lucca

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

Co-Advisor: Prof. Carlos Cueva Herrero, Universidad de Alicante

The dissertation of Juan Francisco Blázquez Pulido has been reviewed by:

Penélope Hernández Rojas, University of Valencia

Marc Vorsatz, National University of Distance Education

IMT School for Advanced Studies Lucca
2026

To Mari Carmen and Juan Antonio.

Contents

List of Figures	xi
List of Tables	xvi
Acknowledgements	xix
Vita and Publications	xxii
Abstract	xxvi
1 Introduction	1
2 Evolution of Conventions in Uncertain Environments	7
2.1 Introduction	7
2.2 The model	11
2.3 Unperturbed dynamics	12
2.4 Perturbed dynamics	13
2.4.1 Uniform mistakes model	15
2.4.2 Scenario/action-dependent mistakes model	18
2.4.3 Condition-dependent mistakes model	21
2.4.4 Cost-dependent mistakes model	23
2.4.5 Scenario pivotality and ergodicity across scenarios	24
2.5 Discussion	25
3 Who's the deceiver? Identifying deceptive intentions in communication	27
3.1 Introduction	27

3.1.1	Related literature	34
3.2	Experimental design and procedure	36
3.2.1	The experiment	36
3.2.2	Identifying senders who aim to deceive	40
3.2.3	Experimental procedure	46
3.3	Results	49
3.3.1	Behavior and beliefs in the strategic game	49
3.3.2	Measuring who aims to deceive in the strategic game	51
3.3.3	The receiver side	60
3.4	Conclusion	61
4	Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment	66
4.1	Introduction	66
4.2	Experimental design and hypotheses	73
4.2.1	The Reporting setting	74
4.2.2	The influence of dishonesty on sanctions	78
4.2.3	The role of truth-telling in sanctions	80
4.2.4	The Mechanisms Behind Sanctions	82
4.2.5	Experimental procedure	84
4.3	Results	89
4.3.1	Punishment for dishonesty	90
4.3.2	The influence of truth-telling on sanctions	92
4.3.3	Discerning lying aversion from truth-telling preferences	94
4.3.4	Norms and Expectations about (Dis)Honesty	97
4.3.5	The role of norms in punishment	104
4.3.6	Senders' behavior	109
4.4	Conclusion	114
5	Conclusion	118
A	Appendix to Chapter 2	122
A.1	Proofs of statements in the main text	122
A.1.1	Proof of Proposition 1	122

A.1.2	Proof of Proposition 2	123
A.1.3	Proof of Proposition 3	124
A.1.4	Proof of Corollary 1	126
A.1.5	Proof of Proposition 4	126
A.1.6	Proof of Proposition 5	129
A.1.7	Proof of Proposition 6	132
A.2	Examples	133
A.2.1	A paradigmatic example	133
A.2.2	Calculations of Example 1	140
A.2.3	Calculations of Example 2	142
B	Appendix to Chapter 3	144
B.1	Analysis of senders' behavior in the strategic game	144
B.2	Comparison of the results in our data with the literature	146
B.3	Robustness analyses ruling out uncertain and incoherent senders	148
B.4	Further analyses to check if random behavior drives our results	150
B.5	Main results controlling by order condition	153
B.6	Experimental procedure diagrams	154
B.7	Sociodemographic characteristics	157
B.8	Instructions	159
B.8.1	Senders' instructions	161
B.8.2	Receivers' instructions	164
C	Appendix to Chapter 4	169
C.1	Supplementary results of punishment and norms	169
C.1.1	Results of messages that do not benefit the sender	169
C.1.2	Norms ECDFs for messages that benefit the sender	174
C.2	Regression results on game-order effects	176
C.2.1	Main regression estimations controlling for game- order effects	177
C.3	Punishment model with situations interacted	179
C.4	Regressions results of norms analyses	182
C.5	Reduced models of punishment and norms	184

C.6	Comparisons norms of observers and judges	185
C.6.1	Comparisons using only the first setting elicited . .	187
C.7	Experimental procedure diagrams	189
C.8	Sociodemographic characteristics	191
C.9	Instructions	192
C.9.1	Observers' instructions	192
C.9.2	Senders' instructions	201
C.9.3	Receivers' instructions	205

List of Figures

1	Decomposition of the behavioral types categorized with the previous method.	43
2	Procedure of the senders who play in the normal order condition.	48
3	Results of the decomposition of the behavioral types with the previous method into the new behaviors using our new method.	54
4	Comparison of the number of senders acting with a deceptive intention captured using the previous method in the literature and our new method.	58
5	Decision page in the Lying Game (harmful situation) from the observer's perspective.	76
6	Decision page in the Choosing Game (harmful situation) from the observer's perspective.	79
7	Decision page in the Truth Game (harmful situation) from the observer's perspective.	81
8	Procedure of observers in the <i>Reporting first</i> game-order condition.	87
9	Punishment across games for senders' payoff-maximizing messages by payoff situation.	90
10	Empirical cumulative distribution functions of punishment for messages that benefit the sender by game and situation.	91

11	Personal, injunctive, and descriptive norms of messages that benefit the sender.	98
12	Distribution of personal norms of sender-benefiting messages.	101
13	Distribution of injunctive norms of sender-benefiting messages.	101
14	Comparison of norms with and without previous punishment elicitation using all observations.	103
15	Frequency of senders' payoff-maximizing reports by game and situation.	110
16	Alluvial plots of senders' reports across games in each situation.	111
17	Alluvial plots of reports across situations for each game. .	113
18	Decomposition of behavioral types under the previous method into those under the new method, for the normal-order condition.	155
19	Decomposition of behavioral types under the previous method into those under the new method, for the inverted-order condition.	156
20	Procedure of the senders who play in the inverted order condition.	156
21	Procedure of the receivers who play in the normal order condition.	157
22	Procedure of the receivers who play in the inverted order condition.	157
23	Page presented to those senders who failed an attention check.	159
24	Experiment introduction page presented to senders.	160
25	Senders' instructions in the strategic task for the normal order condition.	161
26	Senders' page where the belief questions were elicited for the normal order condition.	162

27	Instructions shown to senders between the two games, stating that they are rematched with a different receiver.	162
28	Senders' instructions page in the non-strategic task for the normal order condition.	163
29	Experiment introduction page presented to receivers.	164
30	Receivers' instructions page in the strategic task for the normal order condition.	165
31	Receivers' decision page in the strategic task for the normal order condition.	165
32	Receivers' page where the belief question R_1 in the strategic task was elicited for the normal order condition.	166
33	Receivers' page where the belief question R_2 in the strategic task was elicited for the normal order condition.	166
34	Receivers' page where the belief question R_3 in the strategic task was elicited for the normal order condition.	166
35	Instructions shown to receivers between the two games, stating that they are rematched with a different sender.	167
36	Receivers' instructions page in the non-strategic task for the normal order condition.	167
37	Receivers' decision page in the non-strategic task for the normal order condition.	168
38	Receivers' page where the belief question R_1 in the non-strategic task was elicited for the normal order condition.	168
39	Punishment across games for messages that do not benefit the sender by payoff situation.	170
40	Empirical cumulative distribution functions of punishment for messages that do not benefit the sender by game and situation.	170
41	Personal and Injunctive norms for messages that do not benefit the sender.	171
42	Empirical cumulative distribution functions of personal norms for messages that do not benefit the sender by game and situation.	172

43	Empirical cumulative distribution functions of injunctive norms for messages that do not benefit the sender by game and situation.	172
44	Distribution of personal norms of messages that do not benefit the sender.	173
45	Distribution of injunctive norms of messages that do not benefit the sender.	173
46	Empirical cumulative distribution functions (ECDFs) of personal norms for messages that benefit the sender by game and situation.	174
47	Empirical cumulative distribution functions (ECDFs) of injunctive norms for messages that benefit the sender by game and situation.	174
48	Empirical cumulative distribution functions (ECDFs) of descriptive norms for messages that benefit the sender by game and situation.	175
49	Comparison of norms with and without previous punishment elicitation using observations of the first setting. . . .	187
50	Procedure of senders in the Reporting first game-order condition.	189
51	Procedures for receivers in the between-subject order conditions.	190
52	Instructions of the first part (punishment elicitation) for observers.	192
53	Instructions of the Reporting setting for observers in the punishment part.	193
54	Decision page in the Lying Game from the observer's perspective.	194
55	Decision page in the Truth Game from the observer's perspective.	194
56	Instructions of the Choosing Game for observers in the Reporting-first condition in the punishment part.	195
57	Decision page in the Choosing Game from the observer's perspective.	196

58	Instructions norms part for observers.	196
59	Instructions of the Reporting setting for observers in the norms part.	197
60	Instructions of personal norms for observers.	198
61	Personal norm elicitation in LG from the observer's perspective.	198
62	Instructions of injunctive norms for observers.	199
63	Injunctive norm elicitation in LG from the observer's perspective.	199
64	Instructions of descriptive norms for observers.	200
65	Descriptive norm elicitation in LG from the observer's perspective.	200
66	Introduction of the experiment for senders	201
67	Instructions of the Reporting setting for senders in the Reporting-first condition	202
68	Decision page in the Lying Game from the sender's perspective.	203
69	Decision page in the Truth Game from the sender's perspective	203
70	Instructions of the Choosing game for senders in the Reporting-first condition	204
71	Decision page in the Choosing Game from the sender's perspective	205
72	Experiment introduction for receivers	206
73	Part 1 introduction (norm elicitation) for receivers	206
74	Part 2 introduction (receiver role assigned) for receivers	207
75	Receiver decision page in the reporting setting after getting the sender's message	207
76	Receiver decision page in the Choosing Game after getting the sender's message.	207
77	Instructions of the Reporting setting for receivers	208
78	Instructions of the Choosing Game for receivers	209

List of Tables

1	Senders' behavioral typology using our new method. . . .	41
2	Payoff distribution of each treatment in points.	47
3	Frequency of false messages in the strategic task and senders who expected to be followed by payment treatment. . . .	50
4	Frequencies of the behavioral types categorized with the previous method in the literature in our online sample. . .	51
5	Frequencies of the behavioral types obtained with our new method.	52
6	Payoffs in each situation of the Lying Game, by message veracity.	75
7	Punishment regression results by situation and the pooled sample.	95
8	Correlation matrices for punishment, personal, injunctive, and descriptive norms under alternative samples and levels of aggregation.	105
9	Fixed-effects regressions of punishment on personal, injunctive, and descriptive norms.	106
10	Regressions of punishment on game, message type, and norms.	108
12	Scenario characteristics that are relevant under uniform mistakes.	134

13	Payoff differences between conventions in each scenario that are relevant under intentional mistakes.	136
14	Relevant payoffs in each scenario for condition-dependent mistakes.	137
15	Relevant payoff losses at each scenario under cost-dependent mistakes.	139
16	Scenario characteristics that are relevant under uniform mistakes.	141
17	Payoff differences between convention <i>A</i> and convention <i>B</i> in each scenario.	142
18	Effect of uncertainty in the message sent in the strategic sender-receiver game.	145
19	Comparison of false messages sent in the strategic task between Sutter (2009) and our online subjects data.	146
20	Comparison of the number of senders who expected the receiver to follow the message in the strategic task between Sutter (2009) and our online subjects data.	147
21	Frequencies of the behavioral types obtained using Sutter (2009)'s method.	147
22	Comparison of sender classification under our new identification method across alternative sample restrictions.	149
23	Comparison of the relative number of identified senders with our new method with respect to using the previous method excluding incoherent and senders with S_2 between 40 and 60.	151
24	Comparison of false messages sent in the strategic task by payment treatment in the two order conditions.	153
25	Comparison of the number of senders who expect the receiver to follow the message of the strategic task.	153
26	Frequencies of the behavioral types obtained with our new method controlled by order.	154
27	Sociodemographic characteristics of the sender and receiver participant pools.	158

28	Regression results of the analyses of game-order effects of punishment.	176
29	Fixed-effects regression results controlling for any order effects.	178
30	Differences between payoff situations of the effects of lying and truth-telling on punishment for sender-benefiting messages.	180
31	Results of the regression estimation of Model (C.2).	181
32	Personal norms regressions by situation and for the pooled sample.	182
33	Injunctive norms regressions by situation and for the pooled sample.	183
34	Descriptive norms regressions by situation and for the pooled sample.	183
35	Fixed-effects regressions of punishment on norms: pooled sample.	184
36	Fixed-effects regressions of punishment on norms: harmful situation.	184
37	Fixed-effects regressions of punishment on norms: neutral situation.	185
38	Fixed-effects regressions of punishment on norms: beneficial situation.	185
39	Comparison of norms with and without previously elicited punishment, using all observer decisions.	186
40	Comparison of norms with and without previously elicited punishment, using observations of the first elicited setting.	188
41	Sociodemographic characteristics of the observer, sender, and receiver participant pools.	191

Acknowledgements

These acknowledgments must begin with my parents. This dissertation would never have been possible without your support. Your example has shaped the researcher I am, and being far from you has been one of the hardest parts of this journey.

I extend my deepest gratitude to my supervisors, Prof. Ennio Bilancini and Prof. Carlos Cueva, for their invaluable mentorship throughout my doctorate and for their personal support during the most difficult moments of this doctoral journey. Ennio, thank you for believing in me, introducing me to evolutionary game theory, and rekindling in me the spark to understand human behavior. I will always remember our talks during the first years in your office, where you always offered me a coffee, although I (shamefully) never accepted it. Carlos, I will never forget our (sometimes endless) meetings, our discussions while playing table tennis, and how you broadened my research horizons. I could never have imagined that winning the game in your first game theory class would eventually lead me into behavioral economics. You awakened in me the curiosity and passion for understanding that every researcher needs, and completing this doctorate under your supervision feels like a fitting culmination of this journey.

I am also immensely grateful to Prof. Luca Polonio, my advisor behind the scenes throughout much of this journey. His support goes far beyond his invaluable contribution as a co-author of the second essay (Blazquiz-Pulido, Polonio, and Bilancini, 2024). Thank you, Luca. You taught me that getting inside participants' minds is essential in experimental research. I will never forget our talks over these years. I also thank Prof.

Leonardo Boncinelli for his work as a co-author of the first essay (Bilancini, Blazquiz Pulido, and Boncinelli, 2023).

I would also like to thank Prof. Penélope Hernández and Prof. Marc Vorsatz for their insightful and constructive comments, which helped me improve this dissertation, particularly my Job Market Paper, and encouraged me to consider how the ideas in these chapters could open up future research avenues.

My sincere gratitude also goes to Prof. Marta Serra-Garcia for her guidance during my visit to the Rady School of Management at UCSD. I also owe special thanks to Prof. Ismael Rodríguez-Lara for his support during my visit to the University of Málaga and his suggestions in our meetings thereafter.

I am also deeply grateful to the members of the BEE group at IMT, with special mention to my colleagues Pablo, Tatiana, and Folco, and to my colleagues at the UA, especially Javier. I also extend my gratitude to the members of the FAE department at the UA, with a special mention to Prof. Pedro Albarrán and Prof. Íñigo Iturbe-Ormaeche, always available for advice. I also want to thank Prof. Francesco Serti for his help and advice throughout this joint journey between IMT and the UA.

My heartfelt thanks go to my friends at IMT Lucca, who have become family to me over these years. I am deeply grateful to Alberto, Stefano, Kristina, Chiara, Fra Santucci, Ángel, Nicole, Fra Randone, Víctor, Pablo, and Erica. Their presence was crucial to this journey, standing by me during the hard moments and making the good ones even brighter. *Grazie, ragazzi.*

I also want to thank my lifelong friends from Alicante. To my *matemáticos*, Julen, Santi, Míguel, and Marta. I will always say that studying mathematics was life-changing, the best decision I have ever made, and having met you is one of the main reasons why. Being a mathematician remains an essential part of who I am as a researcher, shaping the way I think and approach problems. Special thanks to Julen and

Santi for making every return home feel easy and for always being there throughout these years. Thanks also to my *Juditas*, especially to Álvaro and Álex, for reminding me that some bonds remain unchanged no matter how much life changes.

Above all, I want to thank Alba, one of the most important people throughout this journey. She has always been my first pilot participant, my confidant through long nights of work, and the person who helped me believe in myself during my weakest moments. Thank you for your love and patience over all these years. None of this would have been possible without you.

Finalmente, reservo este párrafo para mi pilar fundamental, mi familia, pues sin ellos no habría llegado hasta aquí, ni sería quien soy. A mi madre, Mari Carmen, y a mi padre, Juan Antonio, quienes, aunque a veces no pudieran entenderlo, siempre se interesaron por lo que hacía y me apoyaron en todo lo que pudieron y más. Gracias por ser el ejemplo en el que guiarme. A mi hermana, Malena, gracias por tu ayuda cuando no estaba en casa, y por tu apoyo cuando volvía. A mi tía Pili, gracias por ayudarme en mis viajes de un país a otro. Os quiero muchísimo. Por último, quiero dedicar unas palabras a dos personas que, aunque no hayan podido llegar a verlo, sé que estarán orgullosas de mí. A mi tío Jesús, gracias *chache*, siempre fuiste un referente. A mi abuela Magdalena, gracias por tu cariño y apoyo. Te quiero. Sé que me verás acabar *la escuela*, estés donde estés. Por eso, esta tesis va por ti.

Vita

- March 24, 1994** Born, Alicante, Spain.
- 2017** B.Sc. in Mathematics.
Universidad de Alicante, Alicante, Spain.
- 2019** M.Sc. in Quantitative Economics.
Universidad de Alicante, Alicante, Spain.
Thesis: Analyzing sign dependence in financial returns.
Supervisor: *María Ángeles Carnero Fernández*.
- 2019–present** Ph.D. in Economics.
IMT School for Advanced Studies Lucca, Italy.
Double PhD agreement with the University of Alicante.
Supervisor: *Ennio Bilancini*.
Co-supervisor: *Carlos Cueva Herrero*.
- 2023** Visiting PhD student.
Rady School of Management, University of California
San Diego.
San Diego, United States.
Host: *Marta Serra-Garcia*.
- 2024** Visiting PhD student.
Department of Economic Theory and Economic History,
University of Málaga.
Málaga, Spain.
Host: *Ismael Rodríguez-Lara*.

Publications

1. J.F. Blázquez-Pulido, L. Polonio and E. Bilancini, “Who’s the Deceiver? Identifying Deceptive Intentions in Communication,” in *Games and Economic Behavior*, 2024.

Presentations

1. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *6th Meeting of the Behavioral and Experimental Economics Network (VI BEEN 2021)*, University of Bologna, Bologna, Italy, September 2021.
2. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *Young Economists' Meeting 2022 (YEM 2022)*, Masaryk University, Brno, Czech Republic, May 2022.
3. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *Workshop on Strategic Information Transmission (SIT2022)*, University of Alicante, Alicante, Spain, June 2022.
4. "Evolution of Conventions in Uncertain Environments," at the *Learning, Evolution, and Games Conference (LEG2022)*, IMT School for Advanced Studies Lucca, Lucca, Italy, July 2022.
5. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *European Economic Science Association Conference 2022 (ESABOLOGNA 2022)*, University of Bologna, Bologna, Italy, September 2022.
6. "Evolution of Conventions in Uncertain Environments," at the *XVI GRASS Workshop*, Marche Polytechnic University, Ancona, Italy, September 2022.
7. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *Workshop on Model Evaluation and Causal Search: Empirical and Experimental Approaches*, University of Pisa, Pisa, Italy, September 2022.
8. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *47th Symposium of the Spanish Economic Association (SAEe 2022)*, University of Valencia, Valencia, Spain, December 2022.
9. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *12th Southern Europe Experimental Team's Meeting (SEET 2023)*, University of Valencia, Valencia, Spain, February 2023.
10. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *2023 ESA World Meeting (ESAWorld23)*, University of Lyon, Lyon, France, June 2023.
11. "Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment," at the *10th Meeting of the Behavioural and Experimental Economics Network*, University of Trento, Trento, Italy, September 2023.

12. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *64th Symposium of the Italian Economic Association (SIE 2023)*, Gran Sasso Science Institute, L'Aquila, Italy, October 2023.
13. "Who's the Deceiver? Identifying Deceptive Intentions in Communication," at the *48th Symposium of the Spanish Economic Association (SAEe 2023)*, University of Salamanca, Salamanca, Spain, December 2023.
14. "Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment," at the *13th Southern Europe Experimental Team's Meeting (SEET 2024)*, University of Málaga, Málaga, Spain, February 2024.
15. "Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment," at the *Market, Cooperation and Votes Workshop*, Universidad Carlos III, Madrid, Spain, June 2024.
16. "Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment," at the *2nd Milan PhD Economics Workshop*, University of Milan-Bicocca, Milan, Italy, September 2024.
17. "Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment," at the *2nd Scottish Behavioural Sciences Conference 2025*, University of Glasgow, Glasgow, Scotland, May 2025.
18. "Evolution of Conventions in Uncertain Environments," at the *50th Symposium of the Spanish Economic Association (SAEe 2025)*, Universitat Autònoma de Barcelona, Barcelona, Spain, December 2025.
19. "Evolution of Conventions in Uncertain Environments," at the *2026 Biennial conference of the Spanish Royal Mathematical Society (Bienal RSME 2026)*, Universidad de Alicante, Alicante, Spain, January 2026.

Abstract

Many economic interactions are shaped by decisions that depend on others' behavior. This dissertation uses experimental economics and evolutionary game theory to study human behavior across several economic settings, showing how actions may arise from different underlying forces. First, it develops a model of the evolution of conventions under uncertainty, in coordination games where payoffs vary across scenarios following an ergodic Markov process. Using stochastic stability analysis, it shows that rare scenarios may determine long-run equilibrium selection, and that agents' mistakes shape which scenarios become pivotal. Hence, focusing on the most common or average scenario may yield wrong predictions. Then, this dissertation turns to communication, using an experiment to examine behavior in sender–receiver games where truthful messages may be used deceptively. Combining actions, elicited beliefs, and messages in a non-strategic counterpart, it develops a method that identifies more deceptive intentions than previous approaches, distinguishing actual deceivers from pessimistic truth-tellers and detecting senders who conceal their deceptive intentions. Finally, this dissertation studies how third parties punish communication, using an experiment that isolates the roles of a report's veracity and its economic consequences. Results show that sanctions are driven mainly by consequences. Lying plays a secondary and context-dependent role, whereas truth-telling mitigates punishment but does not offset distributional concerns. Besides, sanctions are strongly related to personal and injunctive norms, but not to empirical expectations. Taken together, this dissertation shows that identifying the mechanisms behind observed choices is essential for understanding economic behavior.

Chapter 1

Introduction

«Lo esencial es invisible a los ojos.»

— *Antoine de Saint-Exupéry, El principito*

Communication, coordination, and sanctions are central to many economic interactions. In such environments, individuals often act under uncertainty, transmit private information, and evaluate others' behavior. In economic analysis, it is common to infer preferences, beliefs, or norms directly from observed actions, yet this can be misleading when behavior is shaped by strategic reasoning, moral ambiguity, or reactions to scenarios that rarely occur. Indeed, the same choice may have very different meanings depending on the context in which it is made, and seemingly minor features may have crucial effects on behavior and outcomes.

This dissertation examines three domains using theoretical or experimental approaches: the emergence and persistence of conventions under environmental uncertainty, the identification of deceptive intentions in communication, and the punishment of messages whose veracity and consequences do not necessarily point in the same direction. Across these domains, a common thread emerges: observed choices are often not enough to understand economic behavior, and identifying the mechanisms underlying behavior becomes essential.

Chapter 2 studies the evolution of conventions in environments that change over time. This chapter is based on joint work with Ennio Bilancini and Leonardo Boncinelli, currently available as a working paper on SSRN (Bilancini, Blazquiz Pulido, and Boncinelli, 2023). Much of the literature on evolutionary game theory considers a fixed underlying game to characterize long-run conventions (Foster and P. Young, 1990; H. P. Young, 1993). However, many real-world interactions take place under changing conditions, and conventions that arise under particular circumstances may persist long after those circumstances have ceased to exist.

To capture this feature, this chapter models uncertainty through an ergodic Markov chain. This process governs transitions across coordination games that share a common strategy set but differ in their payoff structures. Using stochastic stability analysis (Ellison, 2000), it shows that long-run convention selection may be determined by one or more scenarios that occur only rarely. As a result, predictions based solely on the most likely scenario or on the average environment can be misleading. The main finding is that the scenarios that matter in the long run depend on how agents make mistakes: which scenario becomes pivotal is determined by the error model governing deviations from best response, even when it occurs with very low probability. Hence, the error structure does not merely determine which equilibria are selected within a given game, but also which scenarios shape long-run behavior.

The main takeaway in this chapter is that rare scenarios and the specification of noise can be decisive for long-run outcomes. This result contributes to the literature on evolutionary game theory, by showing that low-probability scenarios cannot be safely neglected independently of the error model. The chapter also derives tractable expressions to apply stochastic stability analysis under different error models and to identify the conventions preserved in the long run.

Compared to the coordination framework studied in Chapter 2, Chapters 3 and 4 turn to communication settings. Chapter 3 studies deceptive intentions in games of information transmission in which strategic concerns may affect communication. This chapter is based on joint work with

Luca Polonio and Ennio Bilancini, published in *Games and Economic Behavior* (Blazquiz-Pulido, Polonio, and Bilancini, 2024). A central problem in sender-receiver games is that deception is often treated as equivalent to lying (Gneezy, 2005). However, this simplification becomes problematic once strategic considerations are introduced: an individual may tell the truth precisely because she expects the receiver to mistrust her and choose the harmful alternative, a form of deceptive behavior known as *sophisticated truth-telling* (Sutter, 2009). As a result, observing whether a message is true or false is not enough to identify whether the sender intended to deceive. This, in turn, raises a deeper identification problem: how can deceptive motives be measured? Previous approaches have typically relied on the sender’s message and on belief-elicitation questions about the receiver’s response. Chapter 3 shows that this method can misclassify the actual intentions of a substantial fraction of individuals, since senders may also misreport their beliefs in order to conceal their deceptive intentions.

To address this issue, this chapter proposes a new method that combines behavior in a strategic sender-receiver game, answers to a belief-elicitation questionnaire, and behavior in a non-strategic version of the game in which misleading intentions are ruled out. Using this design, it develops a new identification method that captures the motives behind communication more precisely, distinguishing actual deceivers from subjects who only appear deceptive under previous methods.

This new approach uncovers several novel behavioral types previously overlooked in the literature. First, some senders communicate truthfully because they prefer to tell the truth, even though they hold pessimistic beliefs about others’ trustworthiness. These individuals would be erroneously identified as sophisticated deceivers under previous methods. Around 40% of the senders identified as sophisticated truth-tellers with the previous method were actually pessimistic truth-tellers who did not aim to deceive. Second, some senders lie in the game and subsequently offer a (false) justification to rationalize their previous self-interested action in the belief elicitation questionnaire. Around 70% of the senders con-

sidered benevolent liars (i.e., those who lie in order to help the receiver) using the previous method are actually liars who provided an excuse to hide their deceptive intention. Third, some sophisticated truth-tellers hide their deceptive intention once asked about their expectations of the receiver's choice. Around 25% of the senders who appear honest using the previous method are actually suspected of having aimed to deceive through sophisticated truth-telling.

All in all, the new method identifies around 62.5% of the senders as having intended to deceive in the strategic sender-receiver game, which is significantly higher than the 49.1% identified under the previous approach. The main takeaway of this chapter is that deceptive intentions cannot be inferred from observed messages alone, because the same message may reflect different motives depending on the sender's strategic reasoning, and properly identifying those motives becomes crucial. This chapter therefore contributes not only to the literature on dishonesty and strategic communication, but also to the broader study of behavior in economic interactions by showing that human behavior cannot be properly understood from choices and stated beliefs alone, especially when individuals have incentives to conceal the motives behind their actions.

While Chapter 2 studies conventions as behavioral patterns sustained by shared expectations over time, one important mechanism through which they are maintained in practice is punishment: sanctions are often used to deter deviations and reinforce the standards that a community expects its members to follow. At the same time, Chapter 3 shows that communication may conceal deceptive intentions behind seemingly honest behavior, highlighting the difficulty of evaluating which forms of communication should be sanctioned and why. Moreover, in many situations of information transmission, the truthfulness of a message and its consequences do not always point in the same direction. These considerations naturally motivate the questions addressed in Chapter 4: how is communication sanctioned? Is punishment driven by the veracity of a message or by the consequences it entails?

Chapter 4 studies the determinants of punishment in communication

settings. This chapter is based on my single-authored job-market paper. While punishment is one of the main social and institutional tools used to regulate behavior, in games of information transmission it has been mostly studied for its effect on promoting truth-telling rather than for the determinants of punishment itself. But is punishment mainly a response to dishonesty *per se*, reflecting a deontological aversion to lying? Or is it instead driven by the economic consequences of the message?

This chapter addresses these questions using a preregistered experiment in which participants take part in sender-receiver interactions with third-party punishment. In particular, the experimental design varies both dimensions: the message's veracity (using a payoff-equivalent baseline with no lying involved) and the consequences of senders' messages for the receiver (by considering harmful, neutral, and beneficial reports). By exploiting both within-subject variation and between-subject comparisons, this chapter separately quantifies the aggravating role of dishonesty and the mitigating role of truth-telling in punishment.

The main finding is that punishment is primarily driven by the message's economic consequences, especially by whether it benefits the sender at the receiver's expense. Lying plays a secondary and context-dependent role, mattering mostly when consequential concerns are weaker, whereas truth-telling substantially mitigates punishment but does not undo the effect of distributional considerations. Therefore, the two sides of the communication decision have different meanings in the eyes of third-party observers.

Beyond punishment, this chapter studies the normative mechanisms underlying sanctions by eliciting personal norms, injunctive social norms, and empirical expectations about reporting behavior. Sanctions are much more strongly related to personal and injunctive norms than to empirical expectations, suggesting that punishment reflects what individuals believe *should* be done rather than what they expect others to do. In cases of normative disagreement (which arise for morally ambiguous messages such as harmless lies and harmful truths), personal norms are more strongly related to punishment than injunctive norms. Moreover,

controlling for normative appropriateness absorbs much of the aggravating effect of lying, but the effect of economic consequences remains influential, indicating that distributional concerns operate as an additional channel in sanctions beyond normative judgments.

The main takeaway of this chapter is that punishment in communication games cannot be understood solely as a response to dishonesty: what primarily triggers sanctions are the economic consequences of the message and their normative evaluation, with veracity playing a secondary and context-dependent role. This chapter therefore contributes not only to the literature on punishment, communication, and social norms, but also to the broader study of economic behavior by showing that deontological and consequential considerations jointly shape sanctions rather than operating as mutually exclusive forces. In this sense, it complements the previous chapters by showing that even sanctions themselves do not directly reveal the motives behind behavior unless their mechanisms are properly identified.

Taken together, the three chapters highlight the importance of going beyond decisions to understand the mechanisms underlying human behavior. This dissertation contributes not only to the specific areas addressed in each chapter, but also to a broader understanding of economic decision-making by showing that identifying such mechanisms is essential for interpreting individual behavior in economic interactions.

The remainder of this dissertation is organized as follows. Chapter 2 develops a theoretical analysis of convention formation and evolution in uncertain environments. Chapter 3 studies deceptive intentions in sender-receiver games with a strategic component, proposing a new identification method that uncovers behaviors not measured by previous approaches. Chapter 4 examines the punishment of communication, distinguishing the roles of consequences, dishonesty, truth-telling, and norms. Finally, Chapter 5 concludes the dissertation, summarizing its main contributions, discussing the implications of its findings, and outlining directions for future research. Supplementary materials for each chapter are provided in the appendices.

Chapter 2

Evolution of Conventions in Uncertain Environments

This chapter is based on the working paper “*Evolution of Conventions in Uncertain Environments*”, co-authored with Prof. Ennio Bilancini and Prof. Leonardo Boncinelli (Bilancini, Blazquiz Pulido, and Boncinelli, 2023). AI-based tools were used only at a later stage (after the working-paper version had been posted online) and solely for minor editorial assistance, including grammar checking and adapting the text to the thesis template.

2.1 Introduction

A convention may be seen as an established way of coordinating actions in economic and social interactions that have multiple equilibria: once a particular way of doing things becomes established as a rule, it continues in force because participants prefer to conform to the rule given the expectation that others are going to conform as well (Lewis, 2008). Individuals are required to coordinate their actions in many situations, ranging from pure coordination problems (e.g., what side of the road to keep or which platform to adopt) to problems where coordination and conflict co-exist

(e.g., how to share the product of joint work or where to spend time with the partner) and to problems where coordination on superior outcomes is riskier (e.g., what prey to hunt or how much effort to put in teamwork).

The present work is inspired by the recognition that conventions are often established under particular circumstances and then remain in force for a long time, even as those circumstances change. Think for instance of the Covid-19 pandemic, which was thankfully limited in duration, but has shaped conventions in a way that will last for future decades. Often, conventions emerge endogenously with individuals forming expectations on the bases of accumulated evidence and being prone to rare mistakes, suggesting that this process is best modeled as a stochastic dynamical system (H. P. Young, 1996; Sandholm, 2010; Newton, 2018).

A variety of mechanisms that help to sustain conventions has been analyzed in the literature (H. P. Young, 2015). The typical approach is to work with a population of agents who interact in pairs to play a symmetric coordination game, and update behaviors over time in response to past experience and observations. Behaviors in the population usually adjust until convergence is reached to a state where all agents in the population play an equilibrium of the underlying game. For selection purposes, and for the sake of realism, individual mistakes are introduced in the process of behavior updating. This makes some equilibria more prominent than others: rare, mistake-driven transitions may move the system from one convention to another, but some such transitions are much less likely than others. As a result, the system spends most of the time around the convention that is most resilient to mistakes.

This kind of long-run analysis of a perturbed dynamic system is known as stochastic stability, introduced by Foster and P. Young (1990) and then developed by H. P. Young (1993) and Kandori, Mailath, and Rob (1993) building on technical results in Freidlin and Wentzell (1984). Specifically, in this paper, we rely on the radius-coradius theorems in Ely (2002) and Ellison (2000) as introduced in Evans (1993), which are useful tools to simplify the stochastic stability analysis.

In this paper, we consider a two-action symmetric coordination game that is played in a population of boundedly rational agents, who update behaviors according to a myopic best response rule. Importantly, we allow for different *scenarios*, a set of games sharing the same strategy set but each having a specific payoff matrix: at each time the actual scenario can change, bringing about differences in the payoffs of the coordination game. Our idea of having many possible scenarios with different entries in the payoff matrix resembles Markov-switching models in macroeconomics, that allow for parameter values to vary over time, helping to rationalize times series exhibiting sudden changes in the pattern of data (Kim, 1994). After identifying the recurrent classes of the unperturbed Markov chain, we introduce noise in the form of different error models.

In the literature on stochastic stability, it is well known that results strongly depend on the error model: if mutation rates are state-dependent, which equilibrium is selected by stochastic stability depends on exactly how these rates vary from state to state (Bergin and Lipman (1996); see also Blume (2003) for conditions under which the particular noise process does not influence the equilibrium selection results). Therefore, having an appropriate error model turns out to be crucial for stochastic stability to give reliable predictions in terms of equilibrium selection. In this paper, we consider different types of error: uniform (where all mistakes are equally likely, see H. P. Young (1993)), cost-dependent (where more costly mistakes are less likely, see Blume (1993) and, more recently, Alós-Ferrer and Netzer (2010)), scenario/action-dependent (which are a generalization of intentional mistakes, where only actions that can lead to higher payoffs can be chosen by mistake, see Naidu, Hwang, and Bowles (2010)), and condition-dependent (where mistakes are less likely for higher experienced payoffs, see Bilancini and Boncinelli (2020)).

Recent laboratory experiments have shed some light on what error model better fits actual individual behavior. The evidence collected so far suggests that the likelihood of deviations from the myopic best response rule depends negatively both on the past payoff and on expected losses, while intentional mistakes seem to play a more limited role, possibly

concentrated in the early rounds of play (Mäs and Nax, 2016; Lim and Neary, 2016; Hwang, Lim, et al., 2018; Bilancini, Boncinelli, and Nax, 2021). A further issue that has recently attracted interest is the role that risk attitudes can play in the evolution of conventions. One important finding in this respect is that risk aversion is associated with the evolution of maximin conventions, while risk-seeking is associated with the evolution of payoff dominant conventions (Sawa and Wu, 2018; Nax and Newton, 2019). In the current paper, we abstract from risk-attitude considerations.

We find that including a scenario can drastically change the long-run convention selected by stochastic stability, even when the scenario occurs with a very low probability, and therefore one may be tempted to abstract from it. This finding contrasts with Germano (2007), who shows that standard results in evolutionary game theory obtained with a fixed game carry over to uncertain environments where the fixed game is replaced by the average game. The crucial role of extremely unlikely events in our results stems from stochastic stability analysis, which makes some scenarios much more relevant than others, despite the low probability of occurrence: this happens because rare scenarios allow easier transitions through mistakes which become the actual choke points for switching conventions in the long run, as the mistake probability becomes arbitrarily small.

Moreover, which scenario is pivotal in this sense turns out to depend on the error model. This introduces a novel reason why error models are important for stochastic stability analysis, namely the identification of the pivotal scenario (and, hence, the identification of what scenarios one can safely neglect). Finally, we show how to compute the radii and the coradii of conventions for the error models considered in this framework, providing the basis for our main results and facilitating applications to specific cases.

Our contribution raises a general warning for the study of the evolution of conventions: considering the typical or average scenario in which agents have been playing a coordination game over time may lead to neglect the relevant determinants of a convention. Indeed, what happens

under rare and extreme circumstances can determine the ruling convention well after those circumstances have ceased to exist. For a real-world case, consider the Covid-19 emergency. Smart working and contactless payment methods have spread during the Covid-19 emergency and persist even after the emergency is over. Similarly, norm-based behaviors such as avoiding physical contact when greeting people, the use of face-masks, or hand sanitizing have become common practices during the Covid-19 emergency and show considerable persistence.

The remaining of the paper is organized as follows. Section 2.2 introduces the model. Section 2.3 studies the dynamics in the absence of errors, while Section 2.4 considers its perturbed version and collects our main results. Section 2.5 summarizes our contribution and discusses its limitations and possible extensions. All proofs are relegated to Appendix A.1, while Appendix A.2 collects detailed numerical examples with different error models.

2.2 The model

We consider a population of size n , $N = \{1, 2, \dots, n\}$. For simplicity, assume that n is even. Agents interact in a setting described by a finite collection of scenarios, $\mathcal{G} = \{G_1, \dots, G_\varsigma\}$, with $\varsigma > 1$. Each scenario $G_k \in \mathcal{G}$ is a two-player symmetric coordination game with the same action set, $\{A, B\}$, but possibly different payoff structures. For ease of exposition, we refer to $\mathcal{G}$ as *the game*, and to each G_k as a scenario.

Payoffs in scenario G_k are denoted by $\pi_{G_k}^{VZ}$, where $V, Z \in \{A, B\}$ indicate the actions played, respectively, by the recipient and her partner. We assume that, for every $G_k \in \mathcal{G}$, $\pi_{G_k}^{AA} > \pi_{G_k}^{BA}$ and $\pi_{G_k}^{BB} > \pi_{G_k}^{AB}$, so that (A, A) and (B, B) are strict Nash equilibria in each scenario.

Time is discrete and denoted by $t \in \mathbb{N}$. In each period t , a scenario $G_k \in \mathcal{G}$ is realized and agents are randomly matched in pairs to play the coordination game associated with scenario G_k . We denote by $\mathcal{M}$ the set of all possible matchings, that is, all partitions of the set N in pairs.

The current scenario is selected according to the Markov chain $M = (\mathcal{G}, T)$ where T is the transition matrix, with T_{G_l, G_m} indicating the probability that scenario G_m is realized at time t , conditional on scenario G_l having been realized at $t - 1$. The Markov chain M is assumed to be ergodic, that is, every state can be reached from any other state with positive probability, either directly or indirectly.

2.3 Unperturbed dynamics

The population state at time t takes the form $x^t = (\alpha^t, \mu^t, \sigma^t)$, with $\alpha^t = (\alpha_1^t, \dots, \alpha_n^t)$ and $\mu^t = (\mu_1^t, \dots, \mu_n^t) \in \mathcal{M}$, where $\alpha_i^t \in \{A, B\}$ is the action played by agent i at time t , $\mu_i^t \in N \setminus \{i\}$ is the random opponent to whom agent i is matched at time t , and $\sigma^t \in \{G_1, \dots, G_\varsigma\}$ is the scenario where the game is played at time t . So each state describes the actions played by every agent, the partner interacting with every agent and the current scenario. We denote with X the state space.

The revision protocol is as follows. At each t , every agent receives an opportunity to revise her current action with independent probability $\gamma \in (0, 1)$. When doing so, agents adopt a myopic best response rule against the distribution of strategies played at $t - 1$ by the other agents according to the payoff of the current scenario σ^t , randomizing in case of indifference.

We define convention A as the set composed of all states (α, μ, σ) such that for all $i \in N$, $\alpha_i = A$. Analogously, we define convention B as the set composed of all states (α, μ, σ) such that $\alpha_i = B$ for all $i \in N$. Note that this definition of convention is independent of both the realized matching and the realized scenario.

The dynamic system described above induces a Markov chain on X , denoted by $\tilde{M} = (X, \tilde{T})$, where $\tilde{T}$ is the transition matrix with $\tilde{T}_{xx'}$ indicating the probability that state x' is reached at time t , starting from state x at time $t - 1$. By construction, the Markov chain $\tilde{M}$ is consistent with the Markov chain $M = (\mathcal{G}, T)$, for each $G_k \in \mathcal{G}$.

A set of states $W \subseteq X = \{A, B\}^n \times \mathcal{M} \times \mathcal{G}$ is absorbing (or, alternatively, is a recurrent class) according to the induced Markov chain if W is a minimal set with respect to the property that from any $x \in W$ there is zero probability to get, directly or indirectly, to any $x' \in X \setminus W$.

The following proposition establishes what are the absorbing sets in the described unperturbed dynamics.

PROPOSITION 1 (Absorbing sets). *Conventions A and B are the only absorbing sets.*

2.4 Perturbed dynamics

We introduce perturbations to the unperturbed dynamics described in Section 2.3 by allowing for a small probability that agents make mistakes and deviate from the myopic best response. In particular, we assume that the probability that agent i makes a mistake in state $x = (\alpha, \mu, \sigma)$, say by choosing $Z \in \{A, B\}$ that is not best reply at x , approaches zero at the same rate as $\epsilon^{r^Z(i, x)}$ when ϵ goes to zero. We refer to $r^Z(i, x)$ as the resistance of agent i to choose Z at state x , with $Z \in \{A, B\}$. If Z is best reply for agent i at state x , then $r^Z(i, x) = 0$; otherwise, $r^Z(i, x) > 0$. Notice that this formulation allows for mistake rates that depend on the identity of the agent who makes the decision as well as the state at which the decision is made. In the following, we will consider various specifications (error models) of this general formulation.

The introduction of mistakes leads to a perturbed Markov chain on X , denoted by $\tilde{M}^\epsilon = (X, \tilde{T}^\epsilon)$ where $\tilde{T}^\epsilon$ is the perturbed transition matrix obtained from $\tilde{T}$ by adding the perturbations. If $\tilde{T}^\epsilon$ is irreducible, then $\tilde{M}^\epsilon$ admits a unique invariant distribution that describes the fraction of time spent on each state in the long run, irrespectively of the initial state.

The analysis of the perturbed dynamics focuses on the limiting case where ϵ goes to zero, i.e., mistakes become vanishingly small, and the invariant distribution converges to the so-called stochastically stable distribution. A convention is stochastically stable if all its states have positive

probability in the stochastically stable distribution.

To study the stochastically stable distribution in our model, we build upon the techniques developed by H. P. Young (1993) and H. P. Young (1998), which allow to characterize stochastically stable conventions in terms of minimum stochastic potential. In our setting, the *stochastic potential* of convention A is the minimum sum of resistances over paths of states starting at convention B and reaching convention A . The stochastic potential of convention B is obtained analogously. A convention is stochastically stable if its stochastic potential is smaller than or equal to the stochastic potential of the other convention.

Intuitively, the stochastic potential tells us how difficult it is to reach an absorbing set starting from the other absorbing sets: the stochastically stable states are those that are relatively easiest to reach by means of mistakes starting from other states. Hence, the numerical values of resistances, and how they change from transition to transition, are crucial to determine the stochastically stable states.

There are two other notions that we employ in our analysis: the radius and the coradius of a convention, developed by Ellison (2000) and Ely (2002). If E is a union of absorbing sets, its *radius*, denoted by $R(E)$, is the minimum sum of resistances of a path (which is a sequence of states) starting at any state in E and ending in any state of any absorbing set outside E (or, more loosely, the minimum sum of weighted mistakes required to exit E). The *coradius* of E , denoted by $CR(E)$, is the maximum over initial states of the minimum sum of resistances over paths originating from an initial state outside E and ending in a state in E (or, more loosely, the minimum sum of weighted mistakes required to enter E irrespectively of the initial state). Intuitively, $R(E)$ and $CR(E)$ measure how difficult it is to get out of and to get in E , respectively.

By Proposition 1, we know that there are only 2 absorbing sets, i.e., conventions A and B . Using an analogous argument to the one provided in Theorem 1 in Ellison (2000), a convention $Z \in \{A, B\}$ will be stochastically stable if and only if $R(Z) \geq CR(Z)$. Notice that with only two

conventions, the radius of one convention is equal to the coradius of the other convention, and hence the condition can be checked computing radii only. Notice also that, if we consider each single scenario in isolation, we can compute the radius and coradius of each convention (in this case, the states will be composed just by the action profile of the agents and their matches).

In the rest of this section, we characterize radii and coradii of the two conventions for different error models, i.e., distinct specifications of the resistance function $r(i, x)$. Our results allow to establish which convention is stochastically stable depending on the characteristics of the scenarios in $\mathcal{G}$ and the error model considered.

2.4.1 Uniform mistakes model

In the uniform error model, all mistakes are equally likely. Formally, for all $i \in N$ and $x \in X$

$$r^Z(i, x) = r,$$

whenever Z is not the best reply. Specifically, when an agent receives a revision opportunity, with probability $1 - \epsilon$ she will update her strategy by using the myopic best reply rule described in Section 2.3, while with probability ϵ the agent makes a mistake and chooses randomly one strategy in her strategy set. The following proposition characterizes the radius and coradius of the conventions in $\mathcal{G}$ in terms of the radii and coradii of the same conventions considering scenarios $G_1, \dots, G_\zeta$ in isolation. This provides a direct and computationally simple way to establish what convention is selected in the long run.

PROPOSITION 2 (Radius-Coradius calculation for uniform mistakes). *Under the uniform mistakes model, for every convention $Z \in \{A, B\}$:*

1. $R(Z) = \min_k \{R_k(Z)\};$
2. $CR(Z) = \min_k \{CR_k(Z)\},$

where $R_k(Z)$ and $CR_k(Z)$ denote the radius and coradius of convention Z when scenario G_k is considered in isolation.

Proposition 2 shows that, to calculate the radius (coradius) of a convention, it is enough to focus on the scenario whose radius (coradius) is minimal across scenarios, while the actual radii and coradii of the other scenarios are irrelevant. Intuitively, this is because the Markov chain ruling the transition among scenarios is ergodic, so there is always a positive probability of reaching the scenario with the smallest radius (coradius), which therefore provides the path out of (into) the convention with the least resistance.

We observe that Proposition 2 basically provides the minimum number of mistakes needed to leave one convention (for the radius) or to enter into a convention from any possible initial state (for the coradius). There are multiple paths that can lead to this threshold. The uniformity of mistakes implies that mistakes are equally likely in every state. So, the actual path of mistakes done before reaching the threshold number of mistakes (that are needed to get with positive probability to a convention) is irrelevant. This is also important in the coradius calculation: the actual scenario of the initial state turns out not to matter at all since moving to a different scenario comes at no cost in terms of mistakes made by the agents.

To reinforce the intuition of the result in Proposition 2, below we provide an example that illustrates how a scenario that is played very rarely can be crucial in the selection of the long-run convention.

EXAMPLE 1. Consider a game $\check{G}$ composed by the scenarios $\check{G}_1$, $\check{G}_2$ and $\check{G}_3$ described below:

	A	B
A	100,100	0,0
B	0,0	10,10

Scenario $\check{G}_1$

	A	B
A	5,5	-15,3
B	3,-15	6,6

Scenario $\check{G}_2$

	A	B
A	5,5	0,1
B	1,0	5,5

Scenario $\check{G}_3$

with an associated Markov process $\check{M}$ that rules the transitions among $\check{G}_1$, $\check{G}_2$ and $\check{G}_3$ with the following transition matrix:

$$\tilde{T} = \begin{bmatrix} 0.99 & 0.005 & 0.005 \\ 0.99 & 0.01 & 0 \\ 0.99 & 0 & 0.01 \end{bmatrix}, \quad (2.1)$$

so $\tilde{M}$ is ergodic but induces the system to stay approximately 99% of the time in $\check{G}_1$.

In this game, one may be tempted at first glance to think that convention A is the one where the system will spend most time in the long run since $\check{G}_1$ is by far the most likely scenario, and there the basin of attraction of A is ten times the basin of attraction of B .

Let's construct a minimum-resistance path from convention B to convention A : starting at convention B , consider that $\lceil \frac{1}{11}n \rceil$ agents erroneously play¹ action A instead of B , being all placed in any scenario (since under uniform mistakes, errors are equally likely in all scenarios). Then, by the ergodicity of $\tilde{M}$, with positive probability the system moves to scenario $\check{G}_1$ with no mistakes in the process. Hence, at scenario $\check{G}_1$, all agents play action A by best response, reaching convention A with $\lceil \frac{1}{11}n \rceil$ mistakes, each with resistance r . Analogously, a minimum-resistance path can be constructed from convention A to convention B with $\lceil \frac{2}{23}n \rceil$ mistakes in the process, but here the transition to convention B is completed at scenario $\check{G}_2$.

Therefore, applying Proposition 2, we get that $R(B) = \lceil \frac{1}{11}n \rceil r > \lceil \frac{2}{23}n \rceil r = R(A)$. So for n large enough, the only stochastically stable convention is B , contrarily to the idea that could be considered at first sight.

Interestingly, let's consider the reduced game $\check{G}^{(-2)}$, obtained by deleting scenario $\check{G}_2$, and let $\tilde{T}^{(-2)}$ be the associated transition matrix obtained by deleting the second row and column of $\tilde{T}$, and setting $\tilde{T}_{\check{G}_1, \check{G}_1}^{(-2)} = 0.995$. Then, $R(A) = \lceil \frac{4}{9} \rceil > \lceil \frac{1}{11}n \rceil = R(B)$ applying Proposition 2, so convention A turns out to be the unique stochastically stable convention of

¹This is the minimum over scenarios of the minimum number of agents playing action A needed to lead to a transition to convention A in a single scenario with no mistakes in the process, that is, by the rest of agents playing action A by best reply. See Section A.2.2 of the Appendix for all the calculations needed in this example to compute the basins of attraction and the radius of each convention for each scenario.

game $\check{G}^{(-2)}$. However, if we analogously modify the game $\check{G}$ by deleting any other scenario different to $\check{G}_2$, convention B will still be the unique stochastically stable convention. So in this example, scenario $\check{G}_2$ is pivotal for the selection of the long-run convention, and it cannot be neglected despite the low probability that agents can play being placed at it.

2.4.2 Scenario/action-dependent mistakes model

In the scenario/action-dependent mistakes model, the resistance function depends on an agent's action and/or on the current scenario. Formally, for all $i \in N$ and $x \in X$,

$$r^Z(i, x) = r_k^Z,$$

whenever Z is not the best reply, where $Z \in \{A, B\}$ denotes the action chosen by agent i , and $k \in \{1, \dots, \varsigma\}$ denotes the scenario G_k in which the game is played.

The following proposition characterizes the radius and coradius of the conventions in $\mathcal{G}$ in terms of the characteristics of the scenarios $G_1, \dots, G_\varsigma$ considered in isolation. In particular, it turns out that two characteristics of each scenario have to be taken into account: the minimum number of agents playing action Z needed in scenario G_k for action Z to be best reply for the rest of the agents, denoted by $n_k(Z)$ with $Z \in \{A, B\}$, and the resistance to mistakes, r_k^Z , with $Z \in \{A, B\}$. The result below provides a direct and computationally simple way to establish what convention is selected in the long run.

PROPOSITION 3 (Radius-Coradius for scenario/action-dependent mistakes). *Under scenario/action-dependent mistakes:*

1. $R(A) = CR(B) = \min_l \{n_l(B)\} \min_k \{r_k^B\};$
2. $CR(A) = R(B) = \min_l \{n_l(A)\} \min_k \{r_k^A\}.$

Proposition 3 shows that to calculate the radius of convention A , it is enough to focus on the scenario G_l for which $n_l(B)$ is minimal and on the scenario G_k for which r_k^B is minimal, with possibly $G_k \neq G_l$ (and

analogously for the radius of B), while the actual values of these variables for the other scenarios are irrelevant.

Intuitively, this result follows from the ergodicity of the Markov chain ruling the transition among scenarios, which allows the system to transit at no cost (in terms of mistakes) to the scenario where the mistakes have the minimal resistance and then move again at no cost to the scenario where the threshold number of mistakes necessary to trigger a shift to the new convention at no further cost is minimal.

We observe that, differently from what seen for Proposition 2, in Proposition 3 the radius and the coradius are not characterized in terms of the radii and the coradii of the single scenarios considered in isolation, although the relevant characteristics considered are those which allow to calculate these radii and coradii. In particular, while our result allows for $G_k \neq G_l$, relying on the radii and coradii of the single scenarios would impose $G_k = G_l$ and, hence, lead to an overestimation of the radius and the coradius of $\mathcal{G}$.

An interesting particularization of the scenario/action-dependent mistakes model is the case where mistakes are intentional (Naidu, Hwang, and Bowles, 2010; Hwang, Naidu, and Bowles, 2016), that is, mistakes are more likely when they are conducive to a preferred convention with respect to the status quo. In our setting, this particularization is obtained by letting the resistance to mistakes in a given scenario depend on the relative desirability of the two conventions in that scenario.

We consider a straightforward specification in which the likelihood of making a mistake depends positively on the payoff difference between conventions. Formally, for $Z, Z' \in \{A, B\}$ with $Z \neq Z'$, for all $i \in N$, and for all $G_k \in \mathcal{G}$,

$$r^Z(i, x) = r_k^Z \left(\pi_{G_k}^{ZZ} - \pi_{G_k}^{Z'Z'} \right),$$

whenever Z is not the best reply, where r_k^Z is a strictly decreasing function. The characterization of the radius and the coradius under this specification of the intentional mistakes model, presented in the corollary below, follows directly from Proposition 3.

COROLLARY 1 (Radius-Coradius calculation for intentional mistakes).
Under intentional mistakes:

1. $R(A) = CR(B) = \min_k \{r_k^B(\pi_{G_k}^{BB} - \pi_{G_k}^{AA})\} \min_l \{n_l(B)\};$
2. $CR(A) = R(B) = \min_k \{r_k^A(\pi_{G_k}^{AA} - \pi_{G_k}^{BB})\} \min_l \{n_l(A)\}.$

To reinforce the intuition of Proposition 3 and Corollary 1, below we present an example under the intentional mistakes model. In particular, we show how, contrary to the uniform mistakes case, a very rare scenario can be pivotal although its basin of attraction is not crucial in the selection of the long-run convention.

EXAMPLE 2. Consider a game $\hat{G}$ composed by the scenarios $\hat{G}_1, \hat{G}_2$ and $\hat{G}_3$ described below:

	A	B		A	B		A	B
A	10,10	0,0	A	5,5	-15,3	A	3,3	4,0
B	0,0	1,1	B	3,-15	6,6	B	0,4	22,22
Scenario $\hat{G}_1$			Scenario $\hat{G}_2$			Scenario $\hat{G}_3$		

with an associated Markov process $\hat{M}$ that rules the transitions among $\hat{G}_1, \hat{G}_2$ and $\hat{G}_3$ that has the following transition matrix:

$$\hat{T} = \begin{bmatrix} 0.99 & 0.005 & 0.005 \\ 0.99 & 0.01 & 0 \\ 0.99 & 0 & 0.01 \end{bmatrix} \quad (2.2)$$

so $\hat{M}$ is ergodic but induces the system to stay approximately 99% of the time in $\hat{G}_1$.

Similarly to Example 1, at first glance it seems that convention A is the one preserved in the long run, since scenario $\hat{G}_1$ is the most likely scenario, and there the basin of attraction of convention A is far bigger than the basin of attraction of convention B . Moreover, as in the uniform mistakes case, scenario $\hat{G}_2$ is the one in which fewer agents playing action

A are needed² to induce all other agents to switch to action A by best response.

For simplicity, we consider that $r_k^A(z) = r_k^B(z) = \frac{1}{z}$. Then, the scenario in which the resistance to play action B by mistake is lowest is $\dot{G}_3$. Hence, a minimum-resistance path from convention A to convention B is composed by transitions where agents play mistakenly action B occurring only in scenario $\dot{G}_3$, until $\lceil \frac{2}{23}n \rceil$ agents play action B . Then, since $\dot{M}$ is ergodic, with positive probability the system stays at scenario $\dot{G}_2$ with no mistakes in the process, and there convention B is reached with no other mistakes since the rest of agents play action B by best response. Analogously, we can construct a minimum-resistance path from convention B to convention A with $\lceil \frac{1}{11}n \rceil$ mistakes in the process taking all mistakes and the final transition to convention A place in scenario $\dot{G}_1$.

Then, applying Corollary 1, we get that $R(B) = \frac{1}{9} \lceil \frac{1}{11}n \rceil > \frac{1}{19} \lceil \frac{2}{23}n \rceil = R(A)$. So for n large enough, the only stochastically stable convention is B , contrarily to the idea that could be considered at first glance.

However, consider the reduced game $\dot{G}^{(-3)}$, obtained by deleting scenario $\dot{G}_3$, and let $\dot{T}^{(-3)}$ be the associated transition matrix obtained by deleting the second row and column of $\dot{T}$ and setting $\dot{T}_{\dot{G}_1, \dot{G}_1}^{(-3)} = 0.995$. Applying again Corollary 1, $R(A) = \frac{1}{1} \lceil \frac{2}{23}n \rceil > \frac{1}{9} \lceil \frac{1}{11}n \rceil = R(B)$. Hence, convention A becomes the unique stochastically stable convention. So in this example, scenario $\dot{G}_3$ is pivotal for selecting the long-run convention even though the basin of attraction of any convention under this scenario is not fundamental. So this scenario cannot be neglected, even if agents rarely play at it.

2.4.3 Condition-dependent mistakes model

In the condition-dependent mistakes model (Bilancini and Boncinelli, 2020), agents experiencing higher payoffs are less likely to make mistakes.

²See Section A.2.3 of the Appendix for all the relevant features of each scenario for the calculations of radius of each convention, as well as all other calculations needed in this example.

Formally, for all $i \in N$ and $x \in X$,

$$r^Z(i, x) = r(\rho(i, x)),$$

whenever $Z \in \{A, B\}$ is not the best reply, where $\rho(i, x)$ is the actual payoff earned by agent i in state x and r is a strictly increasing function.

The following proposition characterizes the radius and the coradius of the conventions in $\mathcal{G}$ in terms of the characteristics of the scenarios $G_1, \dots, G_\varsigma$ considered in isolation. In particular, the main characteristics that turn out to be relevant are the minimum payoffs that can be earned in each scenario, which in turn maximize the likelihood to make mistakes. The result below provides a direct and computationally simple way to establish what convention is selected in the long run, also numerically once function r is specified.

PROPOSITION 4 (Radius-Coradius calculation for condition-dependent mistakes). *Under condition-dependent mistakes:*

1. $R(A) = CR(B) = \min_k \{r(\pi_{G_k}^{AA})\} + \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\} \min_l \{n_l(B) - 1\}$
2. $CR(A) = R(B) = \min_k \{r(\pi_{G_k}^{BB})\} + \min_k \{\min\{r(\pi_{G_k}^{BA}), r(\pi_{G_k}^{BB})\}\} \min_l \{n_l(A) - 1\}$

Proposition 4 shows that, for a population size n large enough, calculating the radius of convention A requires focusing only on the scenario G_l for which $n_l(B)$ is minimal and on the scenario G_k in which the payoff from playing action A is lowest (and analogously for the radius of B), while the correspondent values all other scenarios are irrelevant. Indeed, when the population is large, the first mistake made along the path to switch from one convention to the other becomes negligible. Moreover, such first mistake is the only one that may be weighted differently from the remaining mistakes in the path, since it is necessarily made in a fully coordinated state, whereas subsequent mistakes may occur after having earned the payoff from miscoordination.

As noted previously for other mistake models, the small number of relevant characteristics needed to compute the radius and coradius is related to the nature of the Markov chain ruling the transition across scenarios. Indeed, the system can transit at no cost (in terms of mistakes) to the scenario where the mistakes have minimal resistance (possibly one scenario for the first mistake and another for the remaining mistakes along the path), and then the system can move again, also at no cost, to the scenario in which the threshold number of mistakes required to trigger a shift to the new convention at no further cost is minimal.

2.4.4 Cost-dependent mistakes model

In the cost-dependent mistakes model, the likelihood of making a mistake is negatively related to the payoff difference between playing the best response and playing the mistaken action. Formally, for all $i \in N$ and $x \in X$,

$$r^Z(i, x) = r_k^Z (\hat{\rho}(i, Z', x) - \hat{\rho}(i, Z, x)),$$

for $Z, Z' \in \{A, B\}$ with $Z' \neq Z$, where Z' denotes the best reply and $\hat{\rho}(i, Z, x)$ is the expected payoff earned by agent i when playing Z in state x .

We characterize the radius and coradius of the conventions in $\mathcal{G}$ in terms of the characteristics of the scenarios $G_1, \dots, G_\varsigma$ considered in isolation. In particular, the main characteristics that turn out to be relevant are the minimum expected losses that can be obtained in each scenario, which in turn maximize the likelihood of mistakes. Since expected losses depend on the current distribution of play, the minimum expected loss for one distribution may arise in one scenario, whereas that for another distribution may arise in a different one. As a result, to get a transition path with minimum total resistance to mistakes it may be needed several shifts across scenarios. The proposition below provides a way to identify such a path and, consequently, how to compute the radius and coradius.

PROPOSITION 5 (Radius-Coradius calculation for cost-dependent mistakes). *Under cost-dependent mistakes:*

1. $R(A) = \sum_{j=0}^{\bar{m}_l(B)} \min_k \left\{ r_k^B \left[\frac{n-1-j}{n-1} (\pi_{G_k}^{AA} - \pi_{G_k}^{BA}) + \frac{j}{n-1} (\pi_{G_k}^{AB} - \pi_{G_k}^{BB}) \right] \right\}.$
2. $CR(A) = \sum_{j=0}^{\bar{m}_l(A)} \min_k \left\{ r_k^A \left[\frac{n-1-j}{n-1} (\pi_{G_k}^{BB} - \pi_{G_k}^{AB}) + \frac{j}{n-1} (\pi_{G_k}^{BA} - \pi_{G_k}^{AA}) \right] \right\}.$

where $\bar{m}_l(Z) = \min_l \{n_l(Z) - 1\}$ for $Z \in \{A, B\}$. In addition, $R(B) = CR(A)$, and $CR(B) = R(A)$.

Proposition 5 shows that, to calculate the radius of convention A , it is enough to focus on the scenario G_k for which $n_k(B)$ is minimal and on a sequence of distinct scenarios $G_{l_1}, \dots, G_{l_h}$, with $h \geq 1$, in which playing B yields the minimum expected loss as the number of agents playing B increases from 0 to $n_k(B) - 1$ (and analogously for the radius of convention B). Intuitively, this result follows from the fact that the expected payoff loss is linear in the fraction of other agents playing B . Hence, if a scenario has provided the minimal expected loss for a given fraction of agents playing B , but it does not provide it anymore for a greater fraction, then it will not provide the minimal expected loss for any larger fraction.

While the formulas in Proposition 5 may appear as computationally demanding and possibly obscure in terms of intuition, in given applications they often boil down to an easy and direct expression that allows establishing which convention is stochastically stable, avoiding cumbersome calculations. We refer the reader to Section A.2.1 of the Appendix, where we illustrate this with an example considering all the error models used in this section.

2.4.5 Scenario pivotality and ergodicity across scenarios

While the exact computation of radius and coradius depends on the error model, a general finding holds for every error model yielding a regular perturbed Markov chain (H. P. Young, 1993, see also Ellison, 2000).

A *regular perturbed Markov chain* is a triple $R = (X, V, (V^\epsilon)_{\epsilon \in (0, \bar{\epsilon})})$, where (X, V) is the unperturbed Markov chain and:

1. (X, V^ϵ) is an ergodic Markov chain, for all $\epsilon \in (0, \bar{\epsilon})$;
2. $\lim_{\epsilon \rightarrow 0} V^\epsilon = V$;
3. There exists a *resistance* function $r : V \times V \rightarrow \mathbb{R}^+ \cup \{\infty\}$ such that for all pairs of states $v, v' \in V$,

$$\begin{cases} \lim_{\epsilon \rightarrow 0} \frac{V_{vv'}^\epsilon}{\epsilon^{r(v, v')}} \text{ exists and is strictly positive} & \text{if } r(v, v') < \infty; \\ V_{vv'}^\epsilon = 0 \text{ for sufficiently small } \epsilon & \text{if } r(v, v') = \infty. \end{cases}$$

Consider a collection of scenario-specific regular perturbed Markov chains, $(R_{G_k})_{G_k \in \mathcal{G}}$, where

$$R_{G_k} = (X, V_{G_k}, (V_{G_k}^\epsilon)_{\epsilon \in (0, \bar{\epsilon})}).$$

Given the Markov chain $M = (\mathcal{G}, T)$ describing the transitions across scenarios, as introduced in Section 2.2, we focus on the associated Markov-switching regular perturbed Markov chain $((R_{G_k})_{G_k \in \mathcal{G}}, \mathcal{G}, T)$.

Let $SS((R_{G_k})_{G_k \in \mathcal{G}}, \mathcal{G}, T)$ denote the set of stochastically stable states. We can then state the following general result.

PROPOSITION 6. *For any pair T' and T'' that are ergodic, the following holds:*

$$SS((R_{G_k})_{G_k \in \mathcal{G}}, \mathcal{G}, T') = SS((R_{G_k})_{G_k \in \mathcal{G}}, \mathcal{G}, T'')$$

This proposition shows that, as long as the scenario process is ergodic, the set of stochastically stable states does not depend on the transition probabilities among scenarios. Intuitively, what matters for long-run selection is whether a scenario can be reached, not how often it occurs. This is why even a very unlikely scenario can still be pivotal.

2.5 Discussion

This paper contributes to the literature on the evolution of conventions by studying their formation and persistence from a novel perspective. In

particular, we show that what happens in rare and extreme circumstances may have long-lasting effects well beyond when those circumstances have ceased to exist. This follows from the fact that, although rare, extreme scenarios can dramatically affect the selection of the long-run convention.

Moreover, we show that the error model governing how agents make mistakes affects which scenario is crucial for the selection of the long-run convention. This introduces a novel reason why error models matter for stochastic stability analysis, namely the identification of which scenarios are pivotal and which can be safely neglected. The intuition behind the main results is illustrated through an example, which is presented in Section A.2.1 of the Appendix for readability.

Our results are obtained under a number of assumptions that limit the scope of the analysis and could be relaxed in future research. First, we assume that agents are myopic best responders. Alternative behavioral rules include imitation (Robson and Vega-Redondo, 1996; Alós-Ferrer and Weidenholzer, 2008), reinforcement learning (Ianni, 2014), logit dynamics (Blume, 1993), and best-experienced payoff dynamics (Sandholm, S. S. Izquierdo, and L. R. Izquierdo, 2019; Sandholm, S. S. Izquierdo, and L. R. Izquierdo, 2020). For broader reviews, see H. P. Young (1998) and Newton (2018). Future work could also allow these behavioral rules to coexist within the same population, along the lines discussed in Newton (2021). Second, our framework could be extended to coordination games with more than two actions. Finally, structured interactions could be considered in place of purely random ones, introducing another important dimension known to affect the evolution of conventions (see Bilancini and Boncinelli, 2020).

All these extensions are likely to complicate the analysis, raising technical issues for deriving tractable expressions for radii and coradii under different error models. At the same time, the main message of this paper remains valid: conventions are often established under specific circumstances that may be both rare and short-lived, yet they can give rise to patterns of behavior that can persist for a long time, as we expect to have been the case during the Covid-19 pandemic.

Chapter 3

Who's the deceiver? Identifying deceptive intentions in communication

This chapter is based on the article “*Who's the deceiver? Identifying deceptive intentions in communication*”, co-authored with Prof. Ennio Bilancini and Prof. Luca Polonio, and published in *Games and Economic Behavior* (Blazquiz-Pulido, Polonio, and Bilancini, 2024). AI-based tools were used only at a later stage (after the article had been published) and solely for minor editorial assistance, including grammar checking and adapting the text to the thesis template.

3.1 Introduction

In many daily-life situations, an individual can exploit information asymmetries to benefit from an uninformed counterpart, such as legal and financial advising, medical recommendations, or political elections. In

such situations, an informed “sender” chooses whether or not to convey a truthful message to an unknowing “receiver”, whose action would impact the payoff to both parties. Hence deception, conceptually defined as a deliberate act to induce a belief that the deceiver considers false in another person, is an important component in the decision-making process of humans when they communicate.

However, deception can happen not only by sending false information but also by telling the truth in the expectation of being mistrusted. An example of the latter behavior is *paltering*, a tactic commonly applied in situations such as negotiation, where the receiver is aware of the possibility of the sender’s deceptive intentions. For instance, let us consider a seller of used cars who would like to persuade a potential client to buy a low-priced car for which she has a large profit margin. The seller could actually tell the truth and say that the cheap car is a lemon, suggesting buying a more expensive car for which the seller’s profit margin is lower. Hence, by telling this truth, the seller could actively try to deceive the buyer believing that the suggestion “do not buy a cheap car because it is a lemon” will be interpreted as a lie. Other real-life scenarios where truth-telling is commonly used deceptively under the expectation of being mistrusted are bluffs in poker or goalkeepers signaling the side of the net where they will dive in a penalty shoot-out.

From an ethical standpoint, one may say that deception is an act that should be avoided, as it has the potential to drive people to make decisions that lead to undesired outcomes. In this regard, it is fundamental to distinguish lying from deception: while lying is the act of claiming something that is not true, deception happens when the claim—irrespective of being true or false—induces a wrong belief in the receiver.

The literature on dishonesty has grown on the evidence that many people do not lie even in situations where it is fully private and beneficial to themselves (Gneezy, 2005). This behavior has been mostly modeled through a pure “lying aversion” component when communicating, that is, assuming that humans experience a disutility—incur a psychological

cost—when telling a lie. Different explanations¹ for lying aversion have been investigated, ranging from intrinsic motivations such as guilt or preferences for doing “the right thing to do” to external factors such as social pressure or image concerns.

Importantly, this literature has focused on lying as the only means of deception, whereas truth-telling was typically considered free of deceptive intentions. An exception to this is Sutter (2009), who documented a potentially different reason for truth-telling in a laboratory experiment: some senders sent the true message in the expectation that the receiver would not follow their advice and then will choose the non-suggested option. Therefore, these apparently benevolent truth-tellers are actually expecting to mislead receivers by telling the truth. This behavior was denoted as *sophisticated truth-telling* or, alternatively, *sophisticated deception*.

To detect such behavior, Sutter (2009) included a belief elicitation questionnaire right after the sender had taken her decision in the sender-receiver game, asking senders for the option they expected the receiver to choose and the probability of their message being followed. By combining the message sent in the game with the option expected to be chosen by the matched receiver, he found that around 30% of the senders aimed to deceive by telling the truth. This result proved that just observing the message emitted is not enough to know the sender’s deceptive intentions. This finding has been considered in posterior studies just to rule out² the possibility that senders use truth-telling as a deceptive method, either by increasing the number of available messages (Erat and Gneezy, 2012) or avoiding it by design (Fischbacher and Föllmi-Heusi, 2013).

This paper aims to clearly identify individuals’ deceptive intentions when communicating in sender-receiver games. To this aim, properly capturing senders’ thoughts while communicating is crucial. However, belief elicitation in this game could be problematic with the sole use of

¹For the sake of exposition, we present further relevant literature on the drivers investigated and how lying aversion is related to different situational or demographic factors in Section 3.1.1.

²A review of the different methods developed to avoid sophisticated deception is provided in Section 3.1.1.

questions about the expected behavior of receivers since the sender may have incentives to hide the deceptive purposes of her message. A common assumption in the lying literature is that individuals have heterogeneous truth-telling preferences. Since the sender-receiver game has a strategic component, the action obtained (i.e., the message sent) depends on the sender's belief about the option chosen by the matched receiver. However, not only do the actions depend on the senders' deceptive intentions, but the elicited beliefs are also influenced by them. Then, although senders' decisions are driven by their beliefs about receivers' actions, solely employing a belief elicitation question right after the game (such as asking about the expected option chosen by the matched receiver) may pose challenges in identifying their deceptive intentions. This is because senders might provide heterogeneous answers to this belief question based on their aim to deceive.

Therefore, a subject with a deceitful intention in the sender-receiver game may have incentives to hide her selfish purposes from the experimenter in the belief elicitation task. This might lead senders to lie when asked about their beliefs in the game, thereby indicating an expectation about the receiver's action that differs from the actual motive behind their own action. On the one hand, this problem of capturing senders' intentions exacerbates with the difference in the incentives between the game and the belief elicitation task, reaching its maximum when beliefs are not incentivized. On the other hand, incentivizing the belief questions does not necessarily solve the problem since some senders may obtain a psychological benefit for keeping their deceptive intentions hidden from the experimenter that is larger than the provided incentive, e.g., if they strongly dislike being considered liars. Note that this problem might appear in any type of communication game where strategic concerns could play a role in the information transmission. To the best of our knowledge, this potential issue has not been previously discussed or considered within the lying literature.

In this paper, we propose a novel method to identify the sender's deceptive intentions for the whole spectrum of senders who aim to deceive

in sender-receiver games with a strategic component, addressing the issues inherent in the previous approach. Specifically, we conducted a two-stage experiment where participants maintained their assigned roles throughout the study. Participants played the classical strategic sender-receiver game, followed by a belief elicitation task about their expectations of the counterpart's behavior. Additionally, they conducted a non-strategic version of the same game where misleading intentions were not allowed. This non-strategic version is equivalent to the classical sender-receiver game with the only difference being that the sender's message directly determines the option chosen for the payments. So in this game, the receiver assumes a passive role, merely receiving the message and observing the final selected option. By combining the messages from the two games, we were able to identify the selfish purposes behind the senders' messages in the strategic sender-receiver game. Then, by using the elicited belief about the expected option chosen by the matched receiver in the strategic game, we accounted for the different intentions that senders held during communication.

To identify senders aiming to deceive in the strategic sender-receiver game and to differentiate them from those who do not, we present evidence of several novel behaviors that a sender might exhibit in such a game, which have not been previously assessed. First, some senders truthfully communicate due to a preference to do so (for example, they might be motivated by moral reasons), although they hold a pessimistic belief about people's trustworthiness. These individuals would be erroneously identified as deceivers if their intentions were solely determined by observing their actions in the strategic game and their beliefs about the expected receiver's choice. Specifically, we found that around 40% of the senders perceived as sophisticated truth-tellers solely using the belief question to discern their deceptive intentions are identified as pessimistic using our new approach, which is grounded in actions and further refined by the belief question.

Second, senders could potentially provide false information during the game and subsequently offer a (false) justification to rationalize their

previous self-interested action in response to the belief elicitation question. These subjects would appear to be benevolent liars (who aim to help the receiver with their message and lack deceptive intentions) when considering solely their actions (sending a false message) and their beliefs regarding the option chosen by the receiver in the strategic sender-receiver game (expressing an expected lack of trust since the receiver will choose the non-suggested option). In contrast, their intention in the game is actually deceptive. We found that around 70% of the senders who would be considered benevolent liars using the previous method in the literature are liars who provided an excuse in these questions to hide their deceiving intention.

Third, there might be senders who sophisticatedly tell the truth to the receiver, expecting not to be trusted, and subsequently hide their deceptive intention once asked about their expectations of the receiver's choice. These subjects would appear to be honest truth-tellers if using the previous method in the literature, whereas their intention in the game was to deceive. We found that around 25% of the senders who would be considered honest using the previous method are actually suspicious of aiming to deceive through sophisticated truth-telling.

Thanks to the identification of these new behavioral types, which were not recognized previously, we can more precisely distinguish between senders who intended to deceive with the message sent in the strategic sender-receiver game and those who did not have the intention to deceive the receiver. Overall, our new method identified that around 62.5% of the senders in our sample aimed to deceive in the strategic sender-receiver game, while only 49.1% of senders would be identified as deceivers using the previous method in the literature.

Although the literature on sender-receiver games is mostly focused on the sender's side,³ the strategic component of this game makes the

³As far as we know, the only contribution that looks into the receiver's trust in sender-receiver games is Gylfason and Olafsdottir (2017). Trustworthiness has been analyzed in other types of settings, such as the distrust game (McEvily, Radzevick, and Weber, 2012) or the investment game (Berg, Dickhaut, and McCabe, 1995), which are less simple to administer than the sender-receiver game and then may trigger different reactions.

receiver's action and beliefs relevant to the senders' decision. Then, exploring beyond the receiver's decision and thoughts can provide insights into the senders' intentions. This is why we study the actions and beliefs of the receivers regarding the messages they get. In particular, we assess trustworthiness by examining how receivers act based on the suggestion provided by the sender. Notably, we observe a significant number of receivers who place trust in the message, regardless of the suggested option.

We also analyze receivers' beliefs about three important aspects of their decision in the game: the truthfulness of the message given, their belief in senders' expectations about their behavior, and whether they considered the message received when deciding their action. Besides, we elicited their belief about the truthfulness of the message also in the game where they are passive. We obtained that their belief about the message's truthfulness is independent of the receiver's role, that is, whether she takes an action or not in the game. Besides, we found that the receivers' expectation regarding the truthfulness of the received message differs from their belief about the sender's expectations of the receivers' trust. Finally, we found that only around 9% of receivers said that they ignored the message in choosing an option. These receivers chose the non-suggested option significantly more than the receivers who considered the message.

The rest of this paper is organized as follows: Subsection 3.1.1 reviews the related literature. Section 3.2 introduces the whole experiment, detailing the different tasks that form it, providing several pre-registered hypotheses, and displaying the details of the design and procedure used to identify senders' deceptive intentions. The experimental results of senders' actions in both games, their elicited beliefs, the behavioral types categorized and the number of senders with a deceptive intention identified using the previous and our new method, together with the results about receivers' trust and their beliefs are presented in Section 3.3. Section 3.4 concludes.

3.1.1 Related literature

Lying aversion has been extensively investigated in the past, both theoretically and empirically. The main models in the literature attribute this behavior to a pure cost of lying (Gneezy, Kajackaite, and Sobel, 2018), guilt aversion (Charness and Dufwenberg, 2006), image concerns (Khalmetzki and Sliwka, 2019), or a combination of them (Kajackaite and Gneezy, 2017). More recently and from a broader perspective, moral preferences (i.e., a preference for doing what is considered “the right thing to do”) have emerged as an explanation for the decisions made by individuals in different interactions (see Capraro, Halpern, and Perc (2024) for a review on this novel topic), including dishonest behavior. An example is the finding by Thielmann, Spadaro, and Balliet (2020) that the honesty-humility personality trait is related to several measures of pro-sociality, suggesting that honesty and different forms of pro-sociality share a common motivation.

Several studies in the dishonesty literature have shown that lying aversion appears to be sensitive to some situational factors, such as the monetary consequences of the lie for both the liar and the affected party (Gneezy, 2005; Hurkens and Kartik, 2009) or the time given to make the decision (Capraro, 2017; Lohse, Simon, and Konrad, 2018; Capraro, Schulz, and D. G. Rand, 2019). Besides, lying aversion seems to be related to demographic characteristics such as gender (Erat and Gneezy, 2012; Capraro, 2018), age (Glätzle-Rützler and Lergetporer, 2015), university studies (López-Pérez and Spiegelman, 2019), and occupation (Besancenot and Vranceanu, 2020).

In the literature on dishonesty, the methods used to study how people aim to deceive have evolved over the years. After discovering that some senders tell the truth with deceptive intentions, many studies in the lying literature have explored different alternatives to rule out potential “strategic beliefs”, thus avoiding the possibility of sophisticated truth-telling. With these changes in the design, they aim to rule out other forms of deception to focus on understanding the factors driving dishonesty.

One approach proposed by Erat and Gneezy (2012) is to expand the message space of the game (from two to, typically, six possible messages). This reduces the expected frequency of receivers following the message that a selfish sender would typically use, making sophisticated truth-telling a more attractive deceptive strategy than straightforward lying. However, Vanberg (2017) showed that in this different version of the game, there are participants who still aim to deceive by telling the truth. Another way to avoid this behavior is to directly rule out any possible strategic concerns in the game by making the receiver passive (Biziou-Van-Pol et al., 2015; Capraro, 2017; Capraro, Schulz, and D. G. Rand, 2019), although this requires a slight change in the framing of the game.

An alternative direction that a branch of the literature has taken is to change the framework completely, moving to the so-called cheating games (Fischbacher and Föllmi-Heusi, 2013; Abeler, Becker, and Falk, 2014; Kajackaite and Gneezy, 2017). However, this variation implies losing the ability to identify dishonest behavior at the individual level, and it usually changes the victim of the lie, with the experimenter becoming a party to the interaction.

Furthermore, from a theoretical perspective, many contributions that model dishonesty explicitly rely on the psychological costs of lying (see Abeler, Nosenzo, and Raymond, 2019; Gneezy, Kajackaite, and Sobel, 2018; Khalmetski and Sliwka, 2019, and the references therein), considering truth-telling as an action free of deceptive intentions. An exception is Sobel (2020), who developed a model to characterize lying and deception in strategic communication settings, differentiating the two concepts. He showed that neither lying needs to be deceptive nor does deception require lying, formally characterizing the differences between both terms. Due to these differences, analyzing just the drivers of lying is not enough to know all the reasons that lead people to try to deceive in sender-receiver games. Then, the above-mentioned variations that the literature followed to avoid sophisticated deception cannot determine the whole spectrum of behaviors that can lead to deception in a sender-receiver game with a strategic component.

3.2 Experimental design and procedure

3.2.1 The experiment

We conducted a preregistered two-stage experiment: Before starting, we randomly assigned to each participant one single role (sender or receiver), which is preserved along the whole experiment. In the first stage, subjects started by playing the conventional sender-receiver game (also called deception game) created by Gneezy (2005), where strategic considerations can play a role (henceforth, the strategic game). Immediately after, senders answered two belief elicitation questions about the receiver's choice designed by Sutter (2009). In the second stage, subjects played a novel sender-receiver game in which no room for strategic considerations is allowed (henceforth, the non-strategic game).

Regarding the participants in the role of the receiver, in addition to their role in the two games, they answered a set of questions right after each of them. In particular, we elicited receivers' beliefs about the truthfulness of the message received in the two games and the senders' expected trustworthiness in the message received in the strategic one. Besides, we asked receivers whether they took into account the message obtained to choose an option in the strategic game. The instructions provided in the introduction of the experiment did not anticipate that more than one sender-receiver game would be played.⁴ This was done to prevent undesired effects on senders' behavior and beliefs that may arise from forward-thinking senders considering the fact that they are playing two games of information transmission.

Notice that some studies have already used a sender-receiver game ruling out any strategic component by allowing senders to choose payments with their messages (Gneezy, Rockenbach, and Serra-Garcia, 2013; Biziou-Van-Pol et al., 2015). However, to our knowledge, this is the first study

⁴Instructions stated: "You will face a sequence of tasks that can involve another randomly picked participant." The full set of instructions used in this experiment is presented in Section B.8 of the Appendix.

with individual data of behavior in both a strategic and non-strategic sender-receiver game, preserving the same setting in both games.

The strategic sender-receiver game

The strategic sender-receiver game is a cheap talk game involving two types of players, the sender⁵ and the receiver, each playing the game once. In this game, two options, labeled A and B, lead to conflicting monetary consequences for both players. The payments associated with each option are known just by the sender, who has to send to the receiver one of the two following predefined messages:

- Message 1: *“Option A will earn you more money than Option B.”*
- Message 2: *“Option B will earn you more money than Option A.”*

Once she gets the message, the receiver, who is fully unaware⁶ of any information about the game’s payoff structure⁷ apart from the message received, must choose which option will determine both players’ earnings.

Moreover, right after emitting the message to the receivers, senders undergo a subsequent belief elicitation task. This task involves answering the same two belief questions created by Sutter (2009), asking about the option they expect the matched receiver to choose (Question S_1) and the fraction of receivers they believe followed the sender’s message (Question S_2). The exact wording of the two questions is:

- Question S_1 : *“Which option do you expect the receiver to choose?”*
- Question S_2 : *“Out of 100 receivers, how many do you think follow the sender’s message on average?”*

⁵To avoid any framing effect and have the best replication as possible of the main related literature, we named the sender as Person 1 and the receiver as Person 2. We keep using the sender and receiver labels throughout the text for the sake of clarity.

⁶Notice that the receiver’s lack of information about the payoffs is also preserved after all decisions are taken. Receivers are only informed about their payoff once they receive their payment, preserving anonymity and hiding the veracity of the sender’s message.

⁷The payoff distribution of each option in our experiment is reported in Table 2, along with other experimental procedure details presented in Section 3.2.3.

To have an experimental design as similar as possible to that of other contributions in the literature, we did not monetarily incentivize the answer to these belief elicitation questions. However, note that monetarily incentivizing the answers would not solve the problem of properly detecting senders' intentions, as explained in Section 3.1.

The standard approach in the literature to detect deceptive intentions in this game was to combine the sender's message with the answer provided to question S_1 . In our new identification method, we adhere to this approach regarding how to use the answers to the belief questions. Specifically, we used the answers to the belief question S_1 in conjunction with additional behavior (as explained in Section 3.2.1) to identify deceptive intentions. As in the standard approach in the literature, the answers to S_2 are reserved for secondary analyses that investigate how the uncertainty in the expected trust affects the message sent in the strategic game.

Unlike most of the literature about sender-receiver games, we elicited receivers' beliefs after making their decision to know the reasons behind trustworthiness. We asked receivers two questions related to their beliefs about the message obtained: the first (Question R_1) refers to the truthfulness of the sender's message, and the second (Question R_2) to their belief about the sender's expectations of receivers following the message obtained, i.e., the second-order belief regarding the behavior answered by the senders to question S_1 . By comparing the two questions, we can provide a measure of how suspicious receivers can be about a possible sophisticated truth emitted by the senders. The last question (Question R_3) aims to know whether receivers considered the information provided by the sender's message to make their decision.

As in the senders' beliefs case, receivers' questions were not monetarily incentivized. The exact wording of the three questions is:

- Question R_1 : *"Out of 100 senders, how many do you think sent a true message?"*
- Question R_2 : *"Out of 100 senders, how many do you think expect the receiver to follow the message sent?"*

- Question R₃: *“Did you take into account the message received to choose an option?”*

The non-strategic sender-receiver game

The non-strategic game is a modified version of the strategic one where the receiver is passive, and the sender’s message implements the option for payments. Except for this, the game remains as in the strategic version. This means that players’ messages and the information provided are exactly as in the strategic game. The single difference is that, although the receiver observes the message emitted by the sender, she takes no action in the game. Notice that the sender was informed that her emitted message and the option selected for the payments in this game would be shown to the receiver.

Therefore, the strategic component that could drive senders’ decision to tell the truth or lie is ruled out in this second game, remaining the factors related to other-regarding preferences and self-interested intentions that senders could have in the strategic game constant. Hence, with this non-strategic version of the game, we can capture senders’ real intentions in the strategic game, ruling out any confound driven by strategic reasoning while controlling for any possible selfish or altruistic motive that could drive senders’ communication. Note that we base our assumption that truth-telling preferences and selfish intentions do not vary across the two games of the experiment both on the full similarity between the two games and several features of our design explained in detail in Section 3.2.3.

Even though they cannot take any action to decide the option selected, receivers observe the sender’s message in the non-strategic game. Hence, we can elicit their beliefs about the truthfulness of the message received to measure how the receiver’s belief about the message’s veracity depends on their role in the game. This question has the same wording as question R₁ in the strategic game, and no monetary incentives were provided. However, as receivers do not make a choice in this game, we cannot ask

analogous questions to R_2 and R_3 .

3.2.2 Identifying senders who aim to deceive

To identify whether a sender aims to deceive the matched receiver (either by telling the truth or lying) in the strategic sender-receiver game, it is necessary to know not only her expectations about the receiver's action, but also the purpose of the action taken. The combination of these three factors drives the behavior that the sender performs in the game.

When Sutter (2009) highlighted the importance of the interaction between the senders' decisions and their beliefs when communicating to receivers, he categorized the behavior of the senders in the classic sender-receiver game into four different behavioral types: benevolent truth-tellers, who told the truth in the game expecting to be followed, plain liars, who lied expecting to be followed, sophisticated truth-tellers, who told the truth deceptively expecting not to be followed, and benevolent liars, who lied expecting not to be followed.

Although Sutter already recognized that other explanations could drive his categories, with our experimental design, we discovered that the previous method conflates senders with different deceptive intentions inside the same category. This implies that the number and the identity of senders considered deceivers in the strategic sender-receiver game could have been partially inaccurate.

To solve this problem, we combine the sender's action (that is, the message sent) in the two (strategic and non-strategic) sender-receiver games with the belief about the sender's expected option chosen by the matched receiver in the strategic game (acquired with the answer to question S_1). This approach offers a more precise method for identifying the intentions of senders in the strategic game, capturing behaviors driven by both deceptive and not deceptive intentions that could not be identified or measured using previous methods.

Table 1 schematizes our new method, providing a description of how

we combine the sender’s message in each of the two sender-receiver games with the sender’s belief about her matched receiver’s expected choice, ending up with eight possible behavioral types. In doing so, we improved the original method, which was solely based on beliefs to identify deceptive intentions in the strategic game, capturing four new behavioral categories.

Sender’s behavioral type with our new method	Str game message	Belief S_1 answer	NS game message	Aim to deceive
Sophisticated truth-teller	Message 1	Option B	Message 2	Yes
Pessimistic truth-teller	Message 1	Option B	Message 1	No
Honest truth-teller	Message 1	Option A	Message 1	No
Incongruous truth-teller	Message 1	Option A	Message 2	Yes
Benevolent liar	Message 2	Option A	Message 1	No
Doubly liar	Message 2	Option A	Message 2	Yes
Plain liar	Message 2	Option B	Message 2	Yes
Incongruous liar	Message 2	Option B	Message 1	Yes

Table 1: Senders’ behavioral typology using our new method based on the veracity of the message sent in the strategic and non-strategic sender-receiver game (message 1 is true whereas message 2 is false), and the expected option chosen by the matched receiver (i.e., the answer to the belief question S_1). The last column remarks the sender’s deceptive intentions for each category.

Our method employs the action in the non-strategic game to account for the monetary-driven preferences that influence the sender’s behavior in the strategic game, ruling out any confounding effect arising from strategic considerations. The last column of Table 1 indicates the behavioral types that aimed to deceive the receiver in the strategic sender-receiver game.

Aligned with our primary objective, we define the concept of *deceptive intention*⁸ in cheap-talk games, serving as the foundation for our new

⁸The difference between this concept and deception is that the latter focuses on the actual

method to identify a sender who aims to deceive.

DEFINITION 1 (Deceptive intention). *A sender has a deceptive intention when transmitting a message to a receiver if, in any case, the aim of her message is leading the receiver to get an inferior outcome.*

Our new method to identify deceptive intentions allows us to distinguish more precisely those senders who aimed to deceive with the message sent from those who did not. The branching of the behavioral types obtained using the previous method to identify deceptive intentions into the behaviors categorized with our new method is presented in Figure 1. The first case involves the denoted sophisticated truth-teller category in the literature, considered a behavior with deceptive intentions. The way to identify this behavior was based on the fact that senders could tell the truth under the belief that the counterpart would not trust their message, then selecting the non-suggested option.

However, there are two types of senders with different motives that could lead to this behavior: on the one hand, the real sophisticated truth-tellers (we preserve the label *sophisticated truth-tellers* when referring to them) who aim to deceive the receiver by telling the truth. These senders would be caught in our non-strategic game when lying to get the monetary benefit.

On the other hand, the second type could be senders primarily motivated by their preferences regarding their actions when communicating. Then, these senders would choose to tell the truth according to their truth-telling preferences. If these players hold a pessimistic belief regarding people's trustworthiness, then their belief about the receiver's trust drives their answer in the belief elicitation question but is not connected with the message they sent in the strategic game. In this case, the player's intention in the strategic game is not to deceive, which is corroborated by their true message in the non-strategic game. We categorize these senders as *pessimistic truth-tellers*. An example of this behavior includes senders

outcome of the interaction, whereas the former deals with the sender's purpose. The two concepts and their definitions have been debated in the literature (see Sobel (2020) for an extensive conceptual and theoretical analysis).

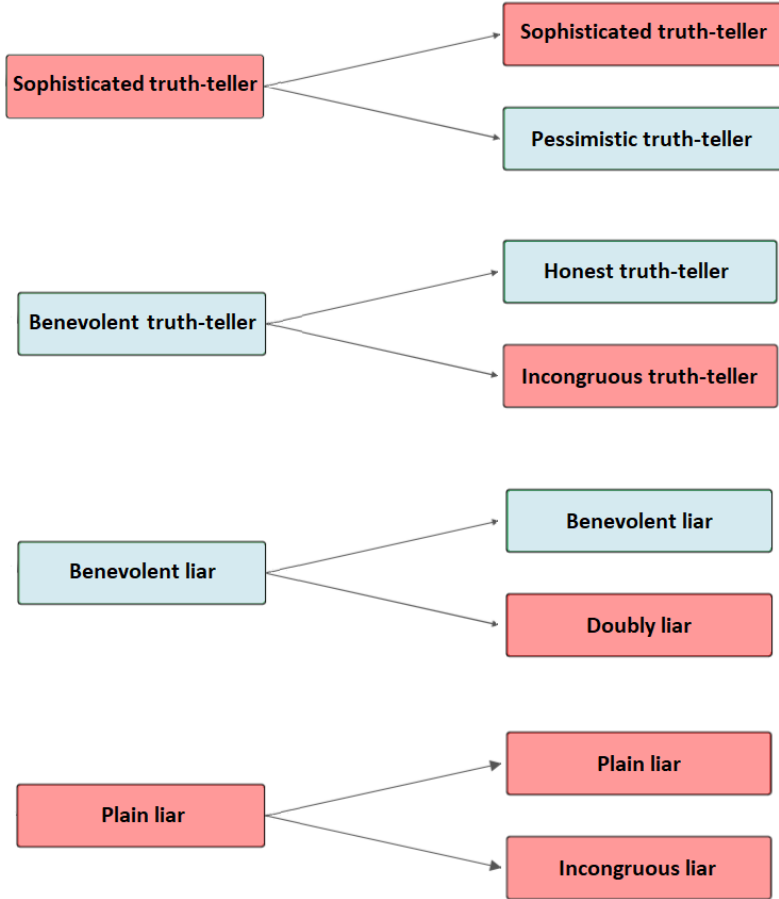


Figure 1: Decomposition of the behavioral types categorized with the previous method that only uses a belief question to identify deceptive intentions in the strategic message (left nodes) into the behaviors categorized with our new method (right nodes) based on actions (both messages) and refined with beliefs. Behaviors in red (blue) are those who (do not) aim to deceive in the strategic game according to each identification method.

who are motivated by moral principles and experience a huge disutility in case they commit a lie, but that also believe people do not place much

trust in others.

Furthermore, to preserve their deceptive intentions hidden, some senders may not express their real thoughts in the belief elicitation questionnaire. This could occur in cases of sophisticated truth-telling, where the sender's self-interested intention becomes even more challenging to identify, or in instances where a sender told a straightforward lie while expecting trust. These senders might subsequently provide false information in the belief questions to justify their prior deceptive behavior. Several reasons, such as social pressure or image concerns about the experimenter's opinion⁹ could explain these two possible behaviors in the strategic game.

We denote as *incongruous truth-tellers* those senders who appear to be benevolent truth-tellers in the strategic game (since they tell the truth claiming to expect to be trusted) but lied when given the opportunity to determine the outcome of the game through their message (as in the non-strategic game).

We hypothesize that the primary behavior underlying this category is that of sophisticated truth-tellers that aimed to conceal their deceptive intentions by providing false information when responding to the belief elicitation question. In the non-strategic game, they are forced to disclose their deceptive behavior to obtain the most beneficial option for themselves. If this selfish reason is the driver of their behavior, then the senders who fall in this category can be considered deceivers. The behavior of these individuals is therefore markedly different from that of *honest truth-tellers*, who genuinely convey the truth in both the strategic and non-strategic games without any intention to deceive the receiver.

⁹Although we specify that anonymity both during and after the whole experiment was preserved, we acknowledge that we cannot reject the possibility that social image plays a role (Jordan and D. G. Rand, 2020). We clearly stated that the Prolific subjects' IDs (the only resource to track participants, although anonymously) would be used just for payments. Besides, running the experiment online in Prolific (whose regulation clearly explains that the experimenters can never identify the participants in the experiments) should be helpful in this concern. A novel result about cheating games in online settings goes in this direction, see Dickinson and McEvoy (2021).

An alternative explanation for this category is that senders could tell the truth in the strategic game driven by a high uncertainty about the receiver's choice since truth-telling preferences have little impact on their decision but are slightly preferred under uncertainty. Then, once they are in the non-strategic game, this uncertainty disappears, allowing them to deceive in order to gain a material benefit.

Our new method identifies individuals who tentatively provided false information in the belief questionnaire, providing an excuse for their observed spiteful behavior (claiming to expect not to be trusted). We denote them as *doubly liars*¹⁰. The deceptive intention of these senders is captured in our non-strategic game, where they once again lie to attain a monetary benefit. Conversely, those *benevolent liars* who aimed not to deceive the receiver (they actually expected receiver's mistrust) consistently tell the truth in the non-strategic game.

Lastly, in our experiment, we can identify senders who deceive in the strategic game stating in the belief elicitation that they expected to be followed, and then subsequently, in the non-strategic game, tell the truth. Although this *incongruous liar* behavior seems odd, this combination could be driven by a change of behavior between the two tasks, meaning that these subjects regret their spiteful behavior in the previous condition. We admit that we cannot identify the drivers of this behavioral type. However, the results presented in Section 3.3 show that this behavior is uncommon, indicating that this category is of minor significance and did not pose an issue in our design.

To summarize, our two-game design allows us to recognize four new behavioral types according to their deceptive intentions that could not be detected just by using the data (action and belief answers) from the strategic game. We formulated several preregistered hypotheses¹¹ about

¹⁰Despite we believe that the main reasoning behind this behavior is the one of lying in the questionnaire to screen a better image and hide their deceptive intentions, another possibility is that it is motivated by a reward gained if telling a lie, meaning that these senders are *lie lovers*. Note that both explanations align in terms of deceptive intentions, thus not affecting our identification process.

¹¹All these hypotheses and analyses, together with other important features of our

the relevance of the new behavioral types categorized with our new method, emphasizing the importance of distinguishing them from the categories previously defined in the literature.

- **Hypothesis 1:** All the new behavioral types categorized with our new method are relevant, that is, none of the new categories is negligible.
- **Hypothesis 2:** The benevolent liar category identified with the previous method in the literature is mainly driven by doubly liars.
- **Hypothesis 3:** A significant number of subjects considered to be sophisticated truth-tellers using the previous method in the literature are pessimistic truth-tellers according to our new method to identify deceptive intentions.
- **Hypothesis 4:** The frequencies of the new behavioral types identified with our new method in our data cannot be fully driven by random behavior.

Crucially, evidence supporting Hypotheses 2 and 3, along with the potential presence of incongruous truth-tellers who are also deceivers, could alter the extent of observed deception in comparison to the results obtained using solely data from the strategic game (action and beliefs), as previously done in the literature.

3.2.3 Experimental procedure

To prevent compensation, reputation, or learning effects, each subject was randomly matched with a different participant for the second interaction. Also, we did not provide any feedback throughout the two tasks, not even the information regarding the first interaction outcome. Besides,

experimental design such as the sample size and the recruitment criteria, were pre-registered on [Aspredicted.org](https://aspredicted.org). The link to the anonymized pre-registration file is <https://aspredicted.org/N1N.KDP>.

anonymity was always preserved throughout the whole experiment, also to the experimenter.

To ensure incentive compatibility and avoid hedging and income effects, the payment of each subject was determined through the random selection of a single game outcome, either based on the payoffs of the option chosen in the strategic game by the receiver or the option selected in the non-strategic game through the sender’s message. The belief elicitation task of the strategic game was not monetarily incentivized, as in Sutter (2009) and other contributions in this literature.

Table 2 presents the three between-subjects treatments we conducted, which differ in the incentives for the sender and the receiver (henceforth payment treatments). These are the ones used in the literature¹² (Gneezy, 2005; Sutter, 2009) to observe how sensitive the sender’s decision is to both players’ relative gains and losses. The two versions of the sender-receiver game (strategic and non-strategic) played by each subject had the same payoff scheme.

Treatment	Option	Sender payoffs	Receiver payoffs	Number of senders
Treatment 1 (T_1)	A	5	6	202
	B	6	5	
Treatment 2 (T_2)	A	5	15	180
	B	6	5	
Treatment 3 (T_3)	A	5	15	211
	B	15	5	

Table 2: Payoff distribution of each treatment in points. Exchange rate: 10 points = 0.5 GBP.

Furthermore, the order of the two tasks (where a task includes the game and its subsequent belief elicitation questionnaire, if any) was randomized among subjects, resulting in two order conditions: some partici-

¹²We maintain the same payoff structure previously adopted in the literature by using experimental points. The exchange rate in this experiment was 10 points = 0.5 GBP.

pants played first the strategic game followed by the subsequent belief elicitation, and then the non-strategic game (henceforth *normal order*), while others played first the non-strategic game and then the strategic game followed by its belief questionnaire (henceforth *inverted order*). Notice that, as explained in Subsection 3.2.1, receivers also answer a belief elicitation question right after the non-strategic game. Before starting the experiment, each participant was randomly assigned to a single role and payment treatment for both games, and an order condition. The diagram in Figure 2 presents the procedure followed by senders who played in the normal order condition. For the procedure of senders in the inverted order and receivers, see Section B.6 of the Appendix.

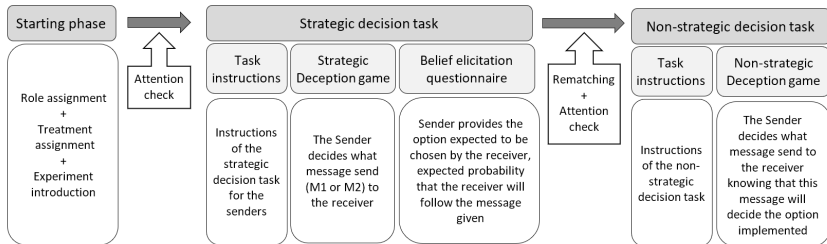


Figure 2: Procedure of the senders who play in the normal order condition.

The experiment was conducted online, from mid-September to early October 2021, recruiting subjects through the experimental platform Prolific (Palan and Schitter, 2018). In total, we recruited 1.186 participants (593 assigned to the role of senders), all US nationals based in the US. 50.3% of the senders and 49.9% of receivers are females and the average age of the subjects is 30.12 (s.d. = 8.17) years old¹³ for senders and 32.12 (s.d. = 8.42) for receivers.

As depicted in Figure 2, an attention check was implemented immediately preceding the instructions of each decision task to ensure the quality of the data collected¹⁴ and to exclude participants who did not pay at-

¹³For more information about the sociodemographic characteristics of the subject pool, see Section B.7 of the Appendix.

¹⁴Peer, Brandimarte, et al. (2017) reported the high quality of the data in Prolific relative

tention. Participants obtained a show-up fee of 0.5 GBP for completing the experiment, and the payoffs of the games were paid through an extra bonus. On average, subjects were paid in total a number of points equal to 7.5 GBP per hour. The experiment took, on average, around 3 minutes to be completed.

3.3 Results

3.3.1 Behavior and beliefs in the strategic game

There are no statistically significant differences between the two order conditions regarding the sender's message transmitted (Pearson's Chi-squared test p-value = 0.22), the expected option chosen by the receiver (Pearson's Chi-squared test p-value = 0.62), and the expectation to be followed by the receiver¹⁵ (Pearson's Chi-squared test p-value = 0.56). This lack of order effects is preserved in all our results. Then, also considering that the order of the two games was randomly assigned, we present the results aggregating order, and if not specified, the strategic game is referred to as the first game, whereas the non-strategic game is also called the second game. For the sake of exposition and to avoid repetitions, the robustness analyses controlling for the order condition are not provided in the main text but presented in brief in Section B.5 of the Appendix.

Table 3 presents, for each payment treatment in our sample, the frequency of false messages sent in the strategic game and whether senders expected to be followed.

As observed in the first row, the payment treatment affects the message sent in the strategic game overall, i.e., T_1 vs. T_2 vs. T_3 (Pearson's Chi-squared test p-value < 0.001), and there is a statistically significant¹⁶

to other alternatives.

¹⁵This value is determined by combining the message sent and the stated belief about the option that the sender expected the receiver to choose.

¹⁶The results of all payment treatment comparisons are presented after correcting for multiple testing using Bonferroni's method. Notice that results are almost identical without

	T₁	T₂	T₃
Send the false message	102/202 (50.5%)	63/180 (35.0%)	127/211 (60.2%)
Expect to be followed	156/202 (77.2%)	144/180 (80.0%)	156/211 (73.9%)

Table 3: Frequency of false messages sent in the strategic task and senders who expected to be followed by payment treatment.

difference between treatments T_1 and T_2 and between T_2 and T_3 (Pearson's Chi-squared test p-value < 0.001), but not between T_1 and T_3 (Pearson's Chi-squared test p-value = 0.18). However, as noticed in the second row, we find no difference among the three payment treatments regarding the expectations to be followed, neither altogether (Pearson's Chi-squared test p-value = 0.36) nor in pairwise comparisons (Pearson's Chi-squared test p-values are 1 for T_1 vs. T_2 , 1 for T_1 vs. T_3 , and 0.586 for T_2 vs. T_3).

The effect of payment treatment on the message sent combined with the lack of significant difference in expected trust between treatments T_2 and T_3 points to the fact that, keeping the cost of lying for the counterpart fixed, the higher the own material relative gain (with respect to the other possible option), the more incentives a sender has to tell a lie. An analogous argument explains the differences found between payment treatments T_1 and T_2 , so keeping the own monetary relative gains fixed, the higher the relative loss for the counterpart (concerning the possible other option), the more difficult it is to tell a lie. A possible explanation for the lack of a significant difference between treatments T_1 and T_3 is that the positive and negative effects observed in the previous pairwise comparisons counterbalance each other due to their opposite directions. This occurs because, even though the sender's relative gains are increased, resulting in a rise in false messages, the receiver's relative losses also increase, causing a reduction in false messages.

correcting for multiple comparisons.

3.3.2 Measuring who aims to deceive in the strategic game

To capture the senders' intentions when communicating in the strategic sender-receiver game, it is necessary to identify the sender's expected trust and how it interacts with the sender's action in the game. First, we provide the results obtained with the method used in the literature, which is based on the interaction between the message sent and the sender's expectation about the receiver's action. As explained in Section 3.2.2, this method relies on the fact that the answers to this question would accurately display the sender's beliefs when transmitting the message, independently of the sender's deceptive intentions. The frequencies of each behavioral type categorized using the previous method in the literature for each payment treatment in our sample are presented in Table 4.

Sender's behavioral type with the previous method	T₁ N=202	T₂ N=180	T₃ N=211	Overall N=593
Benevolent truth-teller	36.1%	55.6%	28.4%	39.3%
Sophisticated truth-teller	13.4%	9.4%	11.4%	11.5%
Plain Liar	41.1%	24.4%	45.5%	37.6%
Benevolent liar	9.4%	10.6%	14.7%	11.6%

Table 4: Frequencies of the behavioral types categorized with the previous method in the literature in our online sample. The first row values (N) display the total number of senders per payment treatment.

We observe payment treatment effects consistent with the results presented in the previous subsection, indicating differences altogether (Fisher's exact test p -value < 0.001) and in all the pairwise comparisons but between T_1 and T_3 (Fisher's exact test p -values are lower than 0.001 for T_1 vs. T_2 , and T_2 vs. T_3 and 0.50 for T_1 vs. T_3). Furthermore, we find no order effects in the results obtained using the previous method in the literature (Fisher's exact test p -value = 0.558). Additionally, we obtained results similar to those in Sutter (2009) regarding the message sent and the sender's expectations to be followed in the strategic game. However, we

observed different frequencies for each behavioral type compared to the previous method in the literature. For the sake of exposition, we refer the reader to Section B.2 of the Appendix for a more detailed comparison of the results obtained in our sample with those from previous contributions in the literature.

The previous results, together with the reasons explained in Subsection 3.2.2, highlight the importance of assessing whether this method correctly discerns senders’ deceptive intentions and whether the previous behavioral types capture all the relevant rationale behind senders’ communication in the strategic game. To reach this goal, we created a new method to identify deceptive intentions, revealing previously unmeasured behavior. The frequencies obtained for all behavioral types categorized with our new method, aggregating and controlling by payment treatments, are presented in Table 5.

Sender’s behavioral type with our new method	T ₁	T ₂	T ₃	Overall
Sophisticated truth-teller	17 (8.4%)	7 (3.9%)	17 (8.1%)	41 (6.9%)
Pessimistic truth-teller	10 (5.0%)	10 (5.6%)	7 (3.3%)	27 (4.6%)
Honest truth-teller	57 (28.2%)	86 (47.8%)	34 (16.1%)	177 (29.8%)
Incongruous truth-teller	16 (7.9%)	14 (7.8%)	26 (12.3%)	56 (9.4%)
Benevolent liar	3 (1.5%)	5 (2.8%)	10 (4.7%)	18 (3.0%)
Doubly liar	16 (7.9%)	14 (7.8%)	21 (10.0%)	51 (8.6%)
Plain liar	68 (33.7%)	36 (20.0%)	83 (39.3%)	187 (31.5%)
Incongruous liar	15 (7.4%)	8 (4.4%)	13 (6.2%)	36 (6.1%)
Total	202 (34.1%)	180 (30.4%)	211 (35.6%)	593

Table 5: Frequencies of the behavioral types obtained with our new method, using the message in the strategic and non-strategic games and the stated expected option chosen by the receiver in the former.

We do not observe any order effects in the frequencies of the categorized behavioral types obtained with our new method, neither aggregat-

ing payment treatments (Fisher’s exact test p-value = 0.27) nor controlling for them (Fisher’s exact test p-values are 0.71 for T_1 , 0.75 for T_2 , and 0.39 for T_3). Besides, none of the frequencies of each category¹⁷ are different across orders. This is why the frequencies in Table 5 and the following results are presented aggregating both order conditions. We also remark that all our results are similar when considering only the normal order condition.

Similar to the previous results regarding payment treatment effects, we found a significant effect both considering all treatments altogether (Fisher’s exact test p-value lower than 0.001) and in all the pairwise comparisons but between T_1 and T_3 (Fisher’s exact test p-values are 0.007 for T_1 vs. T_2 , 0.12 for T_1 vs. T_3 , and lower than 0.001 for T_2 vs. T_3).

The aggregated frequencies obtained show that a significant fraction of senders ($170/593 \simeq 28.7\%$) falls into one of the new behavioral types captured using our method. Besides, these new categories are relevant. To test Hypothesis 1 (none of the new behavioral types is negligible), we used both individual binomial tests and joint Fisher’s exact tests (testing the corresponding hypothesis that they are a fraction below 1% of the data). We consider as the total number of subjects in our analysis either all senders (individual binomial test p-values lower than 0.001, joint Fisher’s exact test p-value = 0.007) or those who fall in one of the new behavioral types identified with our method (individual binomial test p-values lower than 0.001, joint Fisher’s exact tests p-value = 0.015). Besides, we observe significant differences in how the senders are assigned to a new behavioral type between the different categories from the previous method in the literature (joint Fisher’s exact test p-value < 0.001).

As it can be observed by comparing the corresponding frequencies in Tables 4 and 5, the results obtained using our new method vary substantially for the behavioral types identified using the previous approach in the literature. In Figure 3, we decompose¹⁸ the number of senders in

¹⁷The only behavioral type mainly driven by one of the two orders is the incongruous liar. In particular, 24/36 senders who fall in this category played in the inverted order.

¹⁸Note that the percentages in the left balloons are the frequency of senders for each

each behavioral type identified with the previous method in the literature on the number of senders in each behavioral type obtained with our new method.

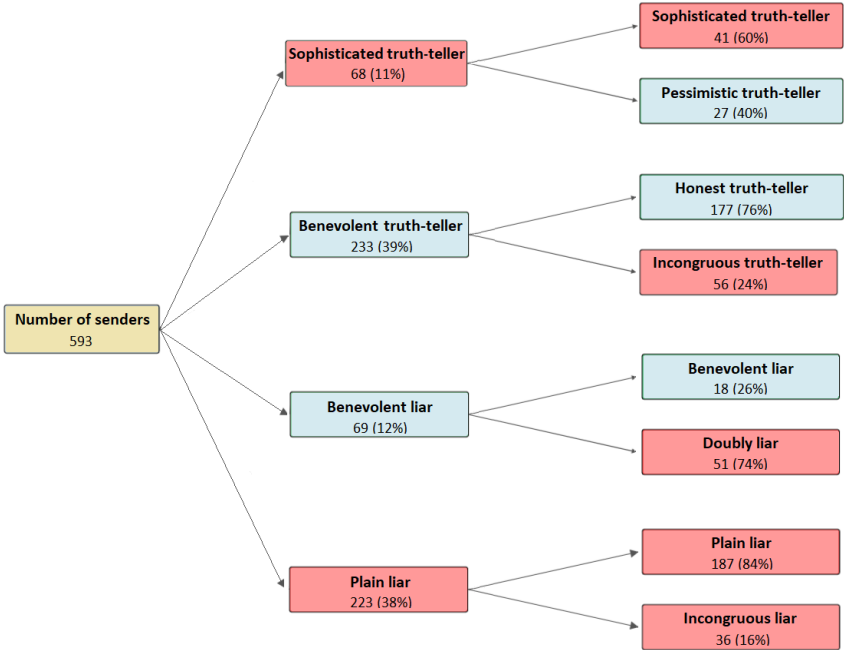


Figure 3: Results of the decomposition of the behavioral types categorized with the previous method into the new behaviors categorized with our new method. Percentages in parentheses are the relative frequencies with respect to the respective parent node. Behaviors in red (blue) are those who (do not) aim to deceive in the strategic game according to each identification method.

Our second hypothesis is also supported by the data: the benevolent liar behavior from the previous method in the literature is mainly driven (51/69, 74%) by senders who probably lied in the belief elicitation

behavioral type categorized with the previous method in the literature over the total number of senders in our sample, whereas the percentages in the right balloons are the frequency of senders of the behavioral types detected with our new method over the ones in the correspondent parent category of the previous method.

questionnaire to excuse the false message sent, in line with Hypothesis 2. Note that this fraction of deceivers would not be identified by just combining the belief elicitation question with the message sent in the strategic sender-receiver game.

We found that a significant number of subjects (27/68, 40%) categorized as sophisticated truth-tellers using the previous method in the literature are just pessimistic, supporting Hypothesis 3. Note that this fraction of senders will be incorrectly identified as deceivers if we use the previous method in the literature to identify deceptive motives.

Moreover, we found that a considerable number (56/233, 24%) of senders designated as benevolent truth-tellers with the previous method in the literature are suspected of being deceivers according to our new method. In particular, they could be sophisticated truth-tellers who, in the belief elicitation question, hid their deceptive intentions of the strategic game by misreporting that they were expecting to be trusted. However, with our experimental design, we would identify the deceptive intentions in their message since they consequently lied in the non-strategic game.

Our further analyses point out that the uncertainty-driven alternative for this behavior proposed in Section 3.2.2 is not driving the behavior of these senders in our data since the sender's uncertainty about the receiver's trust (measured using question S_2) has no significant effect in explaining the message sent in the strategic game. This is analyzed by conducting several regressions to test the role of the sender's uncertainty (modeled using various functions to measure the distance to the full uncertainty point) in explaining the message sent in the strategic game. In none of the regressions¹⁹ conducted, this variable has a significant effect, neither after controlling by the belief about the action of the matched receiver (S_1 , which is statistically significant) nor not doing so.

This evidence suggests that the uncertainty-driven alternative cannot explain our results, and those senders identified as incongruous

¹⁹We used the absolute, the square and the linear function in our regression analyses. Section B.1 of the Appendix contains the detailed results of all these regressions and the distance functions used.

truth-tellers cannot be considered to be plain honest but actually sophisticated truth-tellers who aim to deceive the receiver in the strategic sender-receiver game. It is important to note that, using the previous method in the literature, these senders will be incorrectly identified as subjects not intending to deceive with the message sent.

Furthermore, we conducted additional analyses to examine the role of S_2 in our findings. First, we analyzed how our main results are influenced by senders who expressed complete uncertainty about the receivers' behavior (i.e., $S_2 = 50$), denoted as *fully uncertain*. After discarding them (ending up with 477 senders), there is no statistically significant difference for each behavioral type categorized with our new method (Pearson's Chi-squared test p-value = 0.92). Moreover, the fraction of fully uncertain senders obtained in our sample is not significantly different from the results observed previously in the literature (compared to Sutter (2009), Pearson's Chi-squared test p-value = 0.27). All these analyses are presented in more detail in the Appendix (see Section B.3) together with further checks where we expand the threshold to wider values of S_2 . This provides additional evidence supporting the notion that uncertainty is not a driving factor behind our main results.

In addition, we considered senders as *incoherent* if the combination of their answers to the two belief questions does not align with the message sent in the strategic sender-receiver game. All our main results remain preserved after ruling them out (ending up with 499 senders). Besides, upon excluding both incoherent and fully uncertain senders (ending up with 383 participants), our main findings are preserved or even strengthened, resulting in a higher relative proportion of doubly liars (which is increased to 95% of the number of senders categorized as benevolent liars with the previous method). For the sake of readability, we leave all these robustness analyses to Section B.3 of the Appendix.

Confirming the result obtained using the previous approach in the literature, most senders either plainly tell the truth or lie in both sender-receiver games. However, the fraction of incongruous liars is lower than that of incongruous truth-tellers (proportion test p-value = 0.048). This

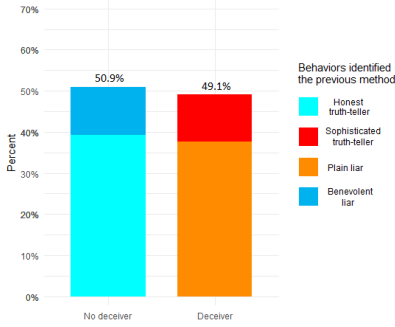
is also in line with our previous arguments that this behavior has little impact on our experimental data. Notice that this behavior is diminished in our robustness analyses after ruling out fully uncertain or incoherent senders. In our view, this points out the lack of relevance of this behavior in our data.

Taking all the previous results together, we compare the fraction of senders that acted with a deceptive intention in the strategic sender-receiver game identified with the previous method in the literature and those using our new method. This comparison is graphically represented in Figure 4. Specifically, Figures 4a and 4d show the fraction of senders with deceptive intentions identified using the previous method in the literature and our new approach. Figure 4c represents the transition between the two methods, displaying how the behavioral types categorized with our new approach would be distributed using the previous method in the literature.

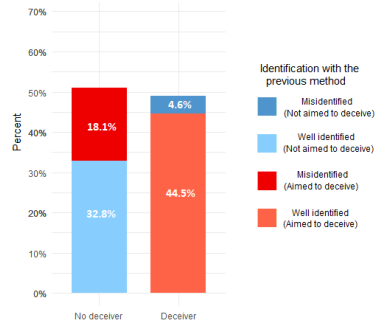
After all, 62.5% (371/593) of the senders in our total sample²⁰ aimed to deceive the receiver according to our new method. This proportion is significantly higher than the 49.1% (291/593) of senders identified using the previous approach in the literature (Pearson's Chi-squared test p -value < 0.001).

This result occurs since the fraction of pessimistic truth-tellers (misidentified as deceivers using the previous approach) is significantly lower than the proportion of senders (both liars or truth-tellers) who are not properly identified with their behavior in the strategic task and whose deceptive intentions are detected with our new approach notwithstanding their misreporting in the belief questions that may hide their deceptive purpose. This can be observed graphically in Figure 4b, which represents the number of senders misidentified (concerning their deceptive intentions) using

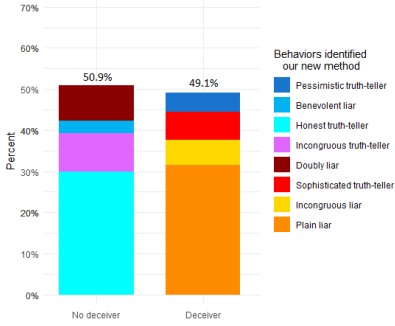
²⁰Even if we observe a trend apparently suggesting that younger participants aim to deceive more than older ones, and that women aim to deceive more than men, we obtained no significant effect of age (t-test p -value = 0.097) or gender (t-test p -value = 0.061) in the fraction of senders identified with a deceptive intention using our new method. We refer the reader to Section B.7 of the Appendix for a more comprehensive study on the effect of age or gender on senders' deceptive intentions.



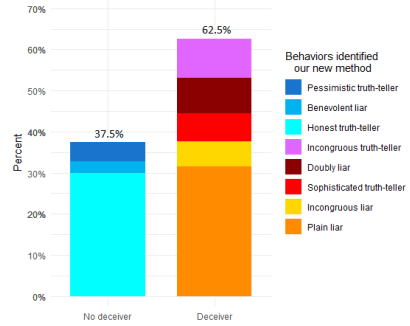
(a) Senders' behavioral types according to the method in the literature (Str. message, S1).



(b) Number of misidentified deceivers and non-deceivers from the previous to our new method.



(c) Distribution of the behavioral types of our new method if identified using the previous method.



(d) Senders' behavioral types according to our new method (Str. message, S1, NS message).

Figure 4: Comparison of the number of senders acting with a deceptive intention captured using the previous method in the literature and our new method.

the previous approach in the literature compared to our new method.

Finally, we conducted several tests to check for the possibility that any possible random behavior²¹ drives the results from the categorization

²¹We are aware that a small amount of random behavior can always be captured in an experiment, and our aim with these analyses is not to state that any player could have been making all her decisions randomly. Our argument is to provide quantitative evidence that, if

obtained using our novel method. First, we checked if the new behavioral types²² proposed were solely driven by random behavior: the results show that the frequencies of the four new behavioral types are jointly different from their respective average (which is equal to 7.17%, proportion test p -value = 0.009). Besides, we can reject the null hypothesis that each new behavioral type is equally likely to happen in our data (Fisher's exact test p -value = 0.003), which would be another potential indication of random behavior.

Moreover, we checked for the possibility that a fraction of senders playing randomly drives the observed results. To do so, we made a battery of tests,²³ one for each noise level (from 1% to 100%) for each of the new behavioral types proposed. We used proportion tests for the hypothesis that a percentage of the senders' is spread uniformly over the four new behavioral types (as it would be if driven by random behavior). We used individual binomial tests for each category to check what levels of randomness could be explained by the data we obtained.

To summarize the results of these randomness tests, we found that there is no level of randomness that can explain the frequencies we captured in the categories obtained with our new method. Specifically, there is no percentage of senders who play randomly that can explain the data in the four new behavioral types, i.e., the intersection of the levels for which we cannot reject the hypothesis of equal distribution across categories is empty.

Therefore, we consider the combination of these two outcomes as an indication that if there is any fraction of senders in the sample who act randomly, their behavior is not driving the results we obtained. Hence, Hypothesis 4 is confirmed, and the new behavioral types proposed are not driven by random behavior.

any, random behavior is not the factor that promotes the results we have already discussed.

²²As a robustness check, we conducted analogous analyses taking all the non-plain categories, that is, all but honest truth-tellers and plain liars. All results are preserved, and the main results are presented in Section B.4 of the Appendix.

²³The results of the whole battery of statistical tests are reported in Section B.4 of the Appendix.

3.3.3 The receiver side

As found in previous contributions, most receivers trusted the message given. In particular, around 87% of the receivers (513/593) chose the option recommended by the sender's message, trusting²⁴ her recommendation.

Another possible explanation for this high rate of receivers who follow the sender's recommendation is that receivers ignore the message obtained, choosing an option without considering the information received. To check if this confound is driving our trust results, we asked receivers about it using an unincentivized question. We found that only around 9% of the receivers (55/593) answered that they did not consider the information obtained. This result provides additional evidence supporting that senders' messages were considered by receivers when deciding, so trust drove receivers' actions. Alternatively, our data shows that receivers took into account the message acquired since, if they did not, they would have randomized the chosen option. However, as it has been shown before, the distribution of their choices is far from random.

Moreover, we elicited receivers' beliefs about the message's truthfulness in the two games. On average, receivers believed that the message obtained was true around two-thirds of the time, both in the strategic and the non-strategic game. Besides, those who did not choose the option suggested in the message believed that the probability of the message being true was lower than the belief of those who followed the message recommendation (Wilcoxon signed-rank test $p\text{-value} < 0.001$).

Comparing the belief in the truthfulness of the message between the two games allows us to examine whether the receivers' role in the game influences their perception of the message's veracity. Truthfulness beliefs in both games were not significantly different (Wilcoxon signed-rank test

²⁴Receivers' trust was independent of the option suggested in the message received (Pearson's Chi-squared test $p\text{-value} = 0.39$). In our experiment, we obtained significantly higher trust rates than previous results in the literature, which ranged around 72% (proportion test $p\text{-value} < 0.001$).

p-value = 0.229). An explanation for why the two games' beliefs are similar is that most receivers just followed the option suggested by the message received without considering its possible falsity since this was the only source of information about the payments they had to make a decision. An interesting avenue for future investigation is to study how the receivers' role in the game affects their beliefs about the message's veracity in settings where receivers have more information than the sender's suggestion.

Lastly, we elicited second-order beliefs from receivers to investigate whether they suspected that senders were employing sophisticated truth-telling behavior. This higher-order reasoning could be a factor influencing senders²⁵ not to use this deceptive method if they anticipated receivers' suspicions.

We obtained that the answers to the first two belief questions (R_1 and R_2) were significantly different (Wilcoxon signed-rank test p-value = 0.003). This result, indicating differences between the belief about the message's truthfulness and the belief about senders' expectations of receivers' trust in the message, could be driven by the fact that some receivers expected non-plain behavior in the strategic game, in particular, sophisticated deception. We cannot provide more information about the drivers of this surprising result, but we consider it an interesting path for the future in order to know better the reasons why receivers trust more than what is expected from an equilibrium viewpoint in sender-receiver games.

3.4 Conclusion

In sender-receiver games where strategic considerations are allowed, senders can deceptively tell the truth to exploit a mistrustful recipient,

²⁵A part of the sender-receiver games literature shows how bounded rationality models such as the level-k reasoning (see Kawagoe and Takizawa (2009) and Wang, Spezio, and Camerer (2010)) or the cognitive hierarchy model (see Li, Özer, and Subramanian (2022)) can explain senders' behavior.

implying that having a deceptive intention does not necessarily coincide with lying. Therefore, the sender's beliefs about the receiver's trustworthiness become crucial for identifying deceptive intentions in these games of information transmission. In this paper, we propose a new method to identify the deceptive intentions of senders in a sender-receiver game characterized by a conflict of interest, where senders may convey truthful information with the intention to deceive. With our new method, we are able to identify four additional behavioral types that would be misclassified in terms of their deceptive intentions using the previous approach in the literature.

We conducted a two-stage experiment composed of two tasks: First, a replica of the classic sender-receiver game in the literature where strategic considerations are allowed, followed by a belief elicitation questionnaire about expected receivers' behavior. Second, an equivalent non-strategic version of the same game with no room for misleading intentions in the message transmitted.

To capture the senders' selfish intentions in the strategic game, we mainly use actions, i.e., the messages sent in the two games. Then, we refine the identification with the answer to the belief question used in the literature, discerning between different deceptive motives. This differs from the previous method, which only uses the belief about the expected receiver's action in the strategic game to infer the sender's intention in the message sent. With our new method, we find that around 40% of the senders who would be considered sophisticated deceivers – using the previous method in the literature – are actually not aiming to deceive with their message. These subjects hold a pessimistic belief about others' trust, being their truth-telling imputed not to a deceptive intention, consistently with their behavior in the non-strategic game.

Our results show that, in games of information transmission where strategic considerations are allowed, relying only on a question about the sender's belief can be insufficient to correctly identify senders' deceptive intentions. The reason is that those senders who aim to deceive in the game, either by telling a plain lie or by sophisticatedly telling the truth,

might have incentives to lie also in a belief elicitation question that reveals to the experimenter their deceptive purpose. Then, these senders might opt for excusing their selfish behavior, preserving their deceptive intentions hidden.

Specifically, we find that around 74% of the senders in our data who appear to be benevolent liars (liars who do not aim to deceive) are actually deceivers who misreported their beliefs to justify the false message sent. Besides, around 25% of the senders who appear to be honest are actually suspicious to have aimed to deceive the receiver by telling the truth sophisticatedly.

We measured how many senders are differently identified (regarding their deceptive intentions) using the previous and our new method. According to our new method, about 62.5% of the senders in our sample aimed to deceive in the strategic sender-receiver game. This proportion is higher than the 49.1% implied by the previous method in the literature. This finding suggests that the single use of a belief elicitation questionnaire right after a sender-receiver game (together with the message's veracity) is not enough to correctly capture the sender's deceptive intentions when strategic considerations are allowed.

An interesting further research step in this regard is to inquire to what extent the heterogeneity implied by our new method is driven by a pure aversion to lying and to what extent it is driven by other sources of benefit from truth-telling, such as moral values or concerns for social image. Another promising avenue for future investigations is to study how the distribution of behavioral types obtained with our new method varies in different communication settings, such as allowing for evasive lying (Khalmetski, Rockenbach, and Werner, 2017) or providing a context with a richer language where partial truths are possible (Alempaki, Doğan, and Saccardo, 2019).

We also provide information about receivers' behavior and beliefs in the sender-receiver game, something little explored before. As in previous contributions, we obtained that receivers had substantial trust in the

senders, and that this trust was independent of the message received. Regarding their stated beliefs, we found that two-thirds of the receivers believe the message obtained is true. Interestingly, we found this belief about the message's truthfulness independent of whether receivers have an active or passive role in the game. An explanation for this result is that the lack of information receivers have in this game might drive this "blind faith". Alternatively, receivers may fail to recognize the strategic and non-strategic nature of the games.

Moreover, around 9% of the receivers ignore the message to choose an option, and those who ignore it mistrust it significantly more. Our receivers' data also indicates that some of them might expect senders to tell the truth while expecting not to be followed. Since some receivers seem to mistrust and others to trust but expect sophisticated deception, we consider an interesting avenue for future research to analyze how the receiver's role and information about the game details affect their trust and beliefs about the message received.

To our knowledge, this is the first study to analyze deception in sender-receiver games where strategic considerations are allowed using an online setting. A few recent contributions provided information about lying behavior in games where no strategic concerns are possible, such as cheating games (Dickinson and McEvoy, 2021) or modified sender-receiver games where the opponent is an external institution that takes no action (Janezic, 2020). However, these games do not allow senders to act strategically, equalizing by design deception to dishonesty.

A word of caution needs to be said about quantifying the spectrum of behaviors concerning deceptive intentions in sender-receiver games. As we noted in Subsection 3.2.2, although we have reasons to think that deceptive intentions are driving those truth-tellers whose stated beliefs are incongruous with their spiteful behavior in the non-strategic game, we cannot fully differentiate those who are aiming to deceive and hiding their intentions by misreporting their beliefs to the experimenter from those who are telling the truth driven by their uncertainty about the receiver's trust.

Another limitation of our study is that the relative size of some categories we detected could depend on the message space, the two-option design, or the conflict of interest between the incentives of senders and receivers in our experiment. Finally, we cannot exclude the possibility that some senders change their preferences according to whether the sender-receiver game has a strategic component or not.

Chapter 4

Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment

This chapter is based on the work “*Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment*” (see Blazquiz-Pulido, 2026). AI-based tools were used only for minor editorial assistance, including grammar checking and adapting the text to the thesis template.

4.1 Introduction

From income tax declarations to financial and medical advice, the transmission of private information plays a central role in many economic and real-world interactions. In such settings, sanctions serve as a key regulatory mechanism to address market failures or maintain trust. It is therefore important to understand how communication is punished in order to explain the mechanisms that sustain information transmission in

society.

Most studies on punishment have focused on strategic or cooperative settings that do not involve communication, where sanctions often serve as a norm-enforcer to deter detrimental behavior and promote socially desirable conduct, although their effects depend on the environment.¹ By contrast, the dishonesty literature has documented the importance of lying aversion (i.e., a disutility some individuals experience from transmitting false information) in shaping reporting decisions (Abeler, Nosenzo, and Raymond, 2019; Gneezy, Kajackaite, and Sobel, 2018), while punishment has mainly been used to study how it promotes truthfulness (Brandts and Charness, 2003; Behnk, Barreda-Tarrazona, and García-Gallego, 2018). However, less is known about what drives such sanctions, and understanding their determinants can shed light on how information transmission is regulated in society.

This paper studies the determinants of punishment in communication games. Previous studies have shown that reporting decisions are influenced both by the message's truthfulness and its payoff consequences (Erat and Gneezy, 2012). However, the impact of these factors on punishment remains largely unexplored. In this paper, we examine the extent to which punishment is driven by the message's veracity, and separate its effect from those of other considerations unrelated to dishonesty, such as the message's economic consequences.

To study this, we conduct an experiment in which participants engage in sender-receiver interactions with third-party punishment. The focus of this paper is on punishment behavior, and to identify the motivations underlying sanctions, we use a communication environment without strategic considerations. Our main setting is a sender-receiver game where the sender privately observes a realized state and has to report the observed outcome to the receiver. We manipulate the reports'

¹For example, punishment has been shown to promote cooperation in the long run (Gächter, Renner, and Sefton, 2008), but it may also backfire by crowding out altruism (Fehr and Rockenbach, 2003) and prosociality (Nosenzo, Xiao, and Xue, 2024). See Xiao (2018) for a review of the effects of punishment on behavior in different settings.

economic consequences so that the sender's payoff-maximizing message may harm, leave unaffected, or benefit the receiver. This allows us to study how punishment responds to the report's veracity while varying the distributional implications of lying and truth-telling.

To isolate the effect of lying, we compare sanctions assigned to false reports with sanctions in a baseline setting in which the same payoff allocations are conveyed without any scope for lying, while preserving the rest of the features of the sender-receiver interaction. These comparisons allow us to measure the effect of dishonesty on sanctions while controlling for any other driving factor, such as distributional or social concerns.

In addition to measuring the effect of lying, our design also allows us to study how truth-telling affects sanctions. By comparing the punishment assigned to selfish truthful reports with that for their payoff-equivalent counterparts, we examine whether being truthful mitigates the punishment for selfish messages. Hence, we distinguish the mitigating role of truth-telling from the aggravating role of lying, assessing if they are similar or instead display asymmetries.

Individuals often use punishment to signal norm violations in a variety of social contexts (Fehr and Fischbacher, 2004; Andrighetto et al., 2013; Kimbrough and Vostroknutov, 2023). Motivated by this expressive function of sanctions, we study whether norms help explain punishment in our sender-receiver setting. In a subsequent part of the experiment, we elicit normative judgments (personal and injunctive social norms) and empirical expectations (descriptive social norms) about reporting behavior. This allows us not only to characterize how different messages are judged, but also to examine how these judgments relate to punishment. Unlike the usual case in the literature, where dishonesty benefits the liar at someone else's expense, in our setting the truthfulness of a report and its economic consequences do not always point in the same direction. This feature is particularly informative from a normative perspective, as it allows us to examine how lying and truth-telling are evaluated across a wider range of situations, including cases in which their appropriateness is not clear-cut.

Our results show that punishment in communication games is primarily driven by the message's economic consequences. Reports that benefit the sender at the receiver's expense get the strongest sanctions. Even when the receiver is unaffected, most participants still punish senders to implement an equal allocation between the two players. Dishonesty increases punishment for harmless messages, although most participants (75%) do not punish a lie that benefits the receiver. However, when the lie harms the receiver, the aggravating effect of dishonesty is only marginal, increasing punishment by about 3.7%.

On the other hand, truth-telling substantially reduces punishment. It lowers sanctions for selfish reports by around 18%, although it does not eliminate them: only about 26% of observers refrain from punishing a harmful truth. Taken together, our findings show that punishment for information transmission is mainly driven by the message's economic implications. Truth-telling and dishonesty have asymmetric effects on sanctions: While truth-telling significantly mitigates punishment for selfish reports, lying plays a secondary context-dependent role and becomes more relevant when consequential concerns are weaker.

Moreover, the effects of lying and truth-telling on norms are broadly consistent with their impact on punishment, with norms additionally responding to lying when dishonesty is harmful. False reports are judged as more inappropriate and are expected to be less frequent than payoff-equivalent non-lying choices, especially when the receiver is harmed. Conversely, truthful reports are evaluated more favorably and are expected to be more frequent. At the same time, the message's payoff implications shape norms, indicating that truthfulness and economic consequences jointly shape how communication is perceived. Besides, the heterogeneous effects of lying on norms show that different types of lies are indeed judged differently: beyond their payoff consequences, lies have a context-dependent effect on normative assessments and beliefs about reporting behavior.

The relationship between norms and punishment shows that sanctions are much more strongly related to personal and injunctive social norms

than to empirical expectations. Reports considered more inappropriate, both personally and socially, receive substantially higher punishment, whereas descriptive beliefs about senders' behavior are weakly related to sanctions. This suggests that punishment reflects what individuals think should be done rather than what they expect others to do. While personal and injunctive norms have a similar relationship with punishment when there is normative agreement about a message's appropriateness, personal norms are more robust when there is no clear consensus.

Overall, differences in normative assessments account for a substantial share of the aggravating effect of lying, which largely disappears once personal and social norms are taken into account. The mitigating effect of truth-telling is also substantially reduced, but remains significant. By contrast, the message's economic consequences preserve a strong influence on punishment even after controlling for its normative appropriateness. These results suggest that, although sanctions are closely related to how messages are judged (especially from a personal perspective), consequential concerns play a substantial role beyond these judgments.

This paper contributes to several strands of the literature. First, it adds to the research on punishment by providing evidence on the determinants of sanctions in games of information transmission. The few existing studies on punishment for dishonesty focus on harmful lies, showing that greater symmetric inequality increases costly punishment and documenting how norm-based nudges affect sanctions (Eisenkopf, Gurtoviy, and Utikal, 2017; Dimant and Gesche, 2023). Our results complement this evidence by showing that although harmful lies are heavily sanctioned, the additional effect of dishonesty *per se* is marginal. What mainly triggers punishment in such cases is the message's harm, rather than the fact that it is false.

Our results on punishment are also relevant for the dishonesty literature. Previous studies document substantial heterogeneity in dishonesty depending on a lie's payoff consequences (Erat and Gneezy, 2012; Janezic, 2020; Babin and Chauhan, 2024), suggesting that lying aversion may itself be context dependent. Our results point to a complementary explana-

tion: some messages may be avoided not only because individuals dislike lying, but also because they are viewed as blameworthy and expected to be sanctioned.² Since senders' payoffs depend on the message sent, reporting decisions are not a clean measure of blameworthiness. By contrast, given its norm-enforcer nature, punishment is an external measure that provides a revealed-preference proxy that complements the reporting evidence in the literature.

To our knowledge, this is the first study to investigate punishment for lies that do not benefit the liar at someone else's expense. These situations are particularly informative because they create a tension between the report's veracity and its social consequences, in contrast to the harmful case, where all considerations point in the same direction, telling the truth. Our results show that lying increases punishment when it is harmless, suggesting that the effect of these lies goes beyond their payoff consequences and therefore has an additional impact on sanctions.

Furthermore, this paper distinguishes the mitigating role of truth-telling from the aggravating effect of dishonesty on punishment, showing that truthfulness can partially legitimize a selfish report. Most of the literature on communication treats truths and lies as opposite sides of the same decision, but our results suggest that their effects on sanctions are asymmetric. We provide new evidence on the distinction proposed by Sánchez-Pagés and Vorsatz (2009) and documented experimentally by Peeters, Vorsatz, and Walzl (2013) and Peeters, Vorsatz, and Walzl (2015), who separate lying aversion from truth-telling preferences by allowing senders to remain silent. To our knowledge, this is the first study to disentangle the roles of dishonesty and truth-telling while preserving the standard message space of communication games.

In addition, this paper contributes to the research on social norms in economic games (Bicchieri, 2005; Kimbrough and Vostroknutov, 2016). In particular, we show that the messages' normative assessments are jointly

²This idea relates to Peeters, Vorsatz, and Walzl (2013), who show that punishers cause excessive truth-telling relative to the equilibrium prediction, both in the presence and in the absence of punishment opportunities.

shaped by their truthfulness and economic consequences, extending the growing work on norms and dishonesty (Alempaki, Doğan, and Y. Yang, 2021; Aycinena et al., 2022; Bicchieri, Dimant, and Sonderegger, 2023; Huber et al., 2023). Besides, we provide evidence from environments in which multiple normative viewpoints coexist (Panizza et al., 2023). Unlike the standard case in economic games with a single clear norm, we show that harmful truths and lies that leave the receiver unaffected do not generate normative consensus, giving rise to conflicting appropriateness judgments.

This paper also contributes to the extensive literature on punishment as a norm enforcer (Axelrod, 1986; Xiao, 2013; Dickinson, Masclet, and Villeval, 2015; Riehm et al., 2022), by showing that sanctions are strongly related to normative assessments, whereas their relationship with empirical expectations is weak. Our findings suggest that when there is no clear consensus about a message's appropriateness, personal norms are more strongly associated with punishment than injunctive norms. This is in line with the findings of Bašić and Verrina (2024), who show that personal norms retain explanatory power beyond injunctive norms and monetary incentives in different economic games.

We also contribute to the recent literature that distinguishes the influence of deontologism from utilitarian considerations in driving behavior (Feess, Kerzenmacher, and Timofeyev, 2022; Bénabou, Falk, and Henkel, 2024). Deontological motives, such as Kantian moral concerns (Bénabou, Falk, Henkel, and Tirole, 2020; Alger and Rivero-Wildemaube, 2024), prohibit certain behaviors based only on the basis of moral rules regardless of their consequences, whereas utilitarian factors, such as distributional and social preferences (Van Leeuwen and Alger, 2024) evaluate decisions based on their consequences. We show that punishment is primarily driven by the economic consequences while truth-telling plays a mitigating role, providing evidence that deontological and consequential concerns interact in the punishment of reports, rather than operating as exclusive forces.

Finally, although the main focus of this study is on punishment,

we also contribute to the experimental work measuring lying aversion (Gneezy, Rockenbach, and Serra-Garcia, 2013; Albertazzi, 2022; Schudy, Grundmann, and Spantig, 2025). Our design identifies lying preferences at the individual level, ruling out confounds driven by strategic beliefs about senders' trustworthiness. Consistent with previous studies, dishonesty increases with the receiver's payoff, suggesting that lying is context dependent. However, around 40% of senders refrain from telling a beneficial lie.³ We measure the context-dependent role of lying and truth-telling (Erat and Gneezy, 2012; Abeler, Nosenzo, and Raymond, 2019), finding that 13% – 38% of senders refrain from making a selfish report when doing so requires dishonesty. Conversely, 34% of senders use a truthful reporting opportunity as a reason to send a harmful message. All in all, 31% of senders never report a false message, consistent with “pure” lying aversion (López-Pérez and Spiegelman, 2013).

The rest of this paper is organized as follows. Section 4.2 describes the experimental design, detailing the different tasks that form it, formalizing our preregistered hypotheses, and explaining the procedure. Section 4.3 presents the experimental results, discussing the main findings. Finally, Section 4.4 concludes.

4.2 Experimental design and hypotheses

We conducted a preregistered experiment in which participants interacted in several games involving three types of players: the sender, the receiver, and the observer.⁴ To prevent any undesired effects on senders' and observers' behavior, the instructions provided in the introduction of the experiment did not anticipate that more than one game would be played, and stated that participants would keep the same role throughout the

³The lying frequencies observed in our setting are not directly comparable to those in studies without punishment, given the well-documented deterrent effect of punishment on dishonesty.

⁴The experimental design (including features such as the sample size and the recruitment criteria), hypotheses and analyses were preregistered on an Open Science Foundation (OSF) project that can be found in this OSF repository link.

experiment.⁵

Although this study primarily focuses on observers' decisions, we first describe each game from the sender's and receiver's perspectives to provide the necessary context, making the observer's role easier to follow. Finally, Section 4.2.5 describes the experimental procedures and logistics, explaining the different order conditions implemented to control for any potential effects driven by the within-subject design.

4.2.1 The Reporting setting

The main setting in this study is a non-strategic sender-receiver game with the possibility of third-party punishment, which we denote the *Reporting setting*. The sender is asked to report a privately observed state to the receiver having the opportunity to lie, while a participant who is not part of this interaction can punish the sender by reducing her earnings depending on the message sent.

Specifically, the sender privately observes a realized state that can take one of two a priori equally likely values (a filled or a striped shape), and is shown the payoffs associated with each possible message she can send. Then, the sender is asked to report the observed state to the receiver, knowing that her report determines both participants' payoffs. Hence, the receiver plays a passive role, taking no action apart from receiving the sender's report.⁶ This ensures that there are no strategic motives in senders' behavior, eliminating any room for beliefs that misinterpret the actual senders' deceptive intentions in sanctioning decisions.

To study the role of dishonesty, we examine the scenario in which

⁵More information about the instructions used in this experiment is provided in Section C.9 of the Appendix. To avoid any framing effect, and consistently with previous studies on dishonesty, we named the sender as Person 1, the receiver as Person 2 and the observer as Person 3. For the sake of clarity, we keep using the sender, receiver, and observer labels throughout the text.

⁶Previous studies found that whether the receiver plays an active or passive role does not significantly affect senders' trustworthiness or their probability of making a false report (Blazquiz-Pulido, Polonio, and Bilancini, 2024).

the sender observes a state that creates a tension between truthfulness and the sender's monetary incentives. In this scenario, a false report always yields a higher payoff for the sender (15 points) than reporting the truth (10 points), whereas the receiver earns 10 points when the truth is reported. For the sake of exposition, we refer to this scenario as the *Lying Game* (LG).⁷

To characterize how punishment for dishonesty depends on the monetary consequences of lying, we elicit decisions in three payoff situations that differ in the economic impact of lying on the receiver: a *harmful* situation, in which lying decreases the receiver's payoff, a *neutral* situation, in which the receiver's payoff is unaffected by the sender's message, and a *beneficial* situation, in which lying increases the receiver's payoff.⁸ We refer to false reports in these situations as harmful, neutral or beneficial lies.⁹ The payoffs associated with each message in each situation of the Lying Game are presented in Table 6. The exchange rate in this experiment was 10 points = 0.5 GBP.

	Sender	Receiver		Sender	Receiver		Sender	Receiver
True	10	10	True	10	10	True	10	10
False	15	5	False	15	10	False	15	15
(a) Harmful			(b) Neutral			(c) Beneficial		

Table 6: Payoffs in each situation of the Lying Game, by message veracity.

The main behavior of interest in this study is that of the observer, who decides whether to assign up to 10 punishment points to the sender for each possible report she can transmit to the receiver. To keep senders' truthfulness unknown to other participants (as is common in the dishon-

⁷An example of the sender's decision page in the Lying Game is shown in Figure 68, provided in Section C.9.2 of the Appendix.

⁸Although each payoff situation could be viewed as a different game, throughout this paper, we use the term *game* to refer to setting variations associated with the state observed by the sender, whereas the term *situation* refers to payoff variations within a given game.

⁹In the dishonesty literature, harmful lies are usually denoted as *black lies*, whereas beneficial lies are denoted as *Pareto white lies*.

esty literature)), we use the strategy method to elicit observer’s decisions.¹⁰ Specifically, conditional on the true outcome, the observer decides how many points (on a scale from 0 to 10) to deduct from the sender’s payoff if she reports each of the two possible messages.¹¹ Figure 5 provides an example of the observer’s decision page for the harmful situation.

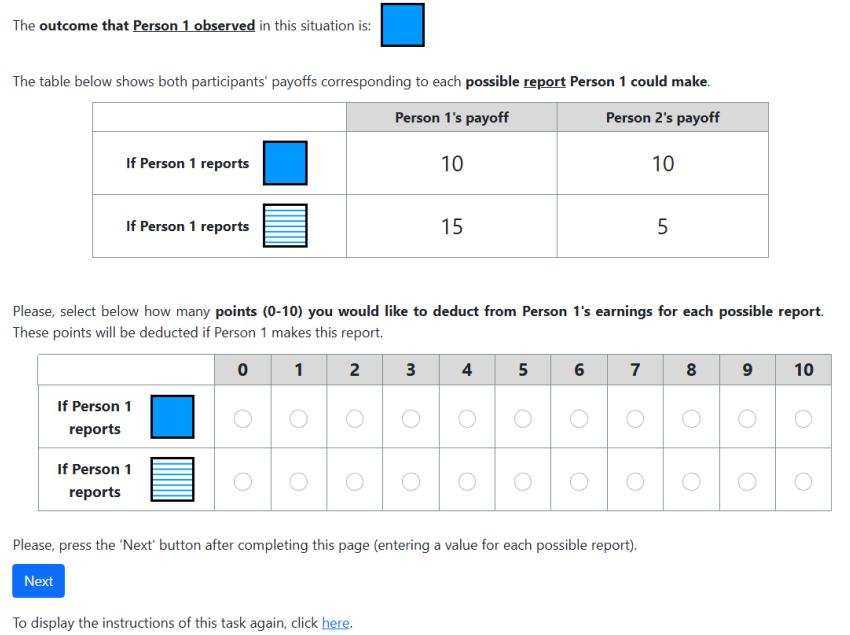


Figure 5: Decision page in the Lying Game (harmful situation) from the observer’s perspective.

Punishment was costless for the observer, but the sender’s earnings were reduced by the number of points assigned to the message sent. We avoided introducing a monetary cost of punishment because doing so would confound punishment decisions, and thus the norm-enforcement nature of sanctions, with participants’ self-interest motives (Cheng et al.,

¹⁰Previous studies show no effect of the strategy method on third-party punishment (Jordan, McAuliffe, and D. Rand, 2016).

¹¹Payoff schemes were designed so that senders’ final game payments cannot be negative.

2022). In particular, observers might reduce sanctions to avoid incurring punishment costs, rather than because the message is viewed as less sanctionable.

Moreover, because sanctions were implemented only for realized reports, the expected cost of punishment would depend on observers' beliefs about the sender's actual behavior. Since our manipulation of the lie's economic consequences may also affect such beliefs, making punishment costly would introduce an additional confounding factor into punishment decisions. This would make it harder to identify whether differences in sanctions are driven by preferences for punishing dishonesty (our object of interest) or by differences in the expected costs of punishment. Although this costless feature may amplify punishment relative to costly sanctions, our main treatment comparisons (described below) net out any inflation that is common across games.

The observer has full information about the game, and both the sender and the receiver are aware of the observer's role. As is common in other games used to measure dishonesty, the receiver is aware of the sender's role and information, but does not know the game's payoff structure. All players are aware of the other players' roles and information.¹²

Therefore, by comparing punishment for false reports across payoff situations, we can study whether sanctions for dishonesty vary with a lie's economic consequences. Building on prior evidence that reporting behavior differs across lie types, we expect punishment for lying to decrease with the receiver's payoff.

Note that unlike the harmful case, in the beneficial situation, consequential concerns point toward lying, whereas a deontological behavior such as "pure" lying aversion (i.e., a dislike for lying regardless of its consequences, see López-Pérez and Spiegelman (2013)) favors truth-telling.¹³

¹²The receiver knows only that payoffs depend on the reported shape, but not exactly how. This information structure is similar to that of the main settings used in the dishonesty literature: the strategic sender-receiver game (Gneezy, 2005) and the cheating game with externalities (Kajackaite and Gneezy, 2017).

¹³In a deontological view, an action is censurable based on its conformity with certain

As a result, a consequentialist view in this situation favors the opposite action to deontological reasoning (Feess, Kerzenmacher, and Timofeyev, 2022; Bénabou, Falk, and Henkel, 2024), and then judgments of dishonesty are less straightforward. This tension may affect sanctions, potentially making truth-telling punishable for a non-negligible fraction of observers.

4.2.2 The influence of dishonesty on sanctions

In the Reporting setting, senders' reports affect other's payoffs. Hence, sanctions for dishonesty may also reflect factors that are not related to lying *per se*, such as distributional preferences, efficiency concerns, or a general willingness to punish.

To isolate the effect of lying, we use a baseline setting with no lying involved that preserves the communication structure of the Reporting setting. In this counterfactual scenario, denoted the *Choosing Game* (CG), the sender has to choose one of two possible alternatives, knowing the payoff allocations by each choice, and transmits their choice to the receiver. The rest of the game's features are as in the Reporting setting, so the receiver is unaware of the game's payoff structure, and the information sets of all roles are preserved.

As in the Lying Game, we elicit senders' and observers' decisions in three payoff situations. Consistent with the notation for lie types, we refer to the corresponding actions in the Choosing Game as harmful, neutral, and beneficial choices. An example of the sender's decision page in the Choosing Game is shown in Figure 71 in Appendix C.9.2, and Figure 6 shows an example of the observer's decision page in the harmful situation.

In this game, observers decide in a setting that mirrors the Lying Game in all respects, except that the sender's action does not involve dishonesty.¹⁴ As a result, there is no scope for lying preferences to affect

rules, regardless of its consequences (Kant, 1785; Alexander and Moore, 2007).

¹⁴All framing features in the game, such as the shapes, patterns, and colors used to

The table below shows both participants' payoffs corresponding to each possible choice Person 1 could make.

	Person 1's payoff	Person 2's payoff
If Person 1 chooses 	10	10
If Person 1 chooses 	15	5

Please, select below how many points (0-10) you would like to deduct from Person 1's earnings for each possible choice. These points will be deducted if Person 1 makes this choice.

	0	1	2	3	4	5	6	7	8	9	10
If Person 1 chooses 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
If Person 1 chooses 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Please, press the 'Next' button after completing this page (entering a value for each possible choice).

Next

To display the instructions of this task again, click [here](#).

Figure 6: Decision page in the Choosing Game (harmful situation) from the observer's perspective.

observers’ decisions in this game, while other preferences that may drive punishment are preserved.

By comparing punishment for false reports in the Lying Game with punishment for the corresponding choices in the Choosing Game, we can measure the extent to which pure lying preferences (driven by deontological factors) influence sanctions for dishonesty relative to the report’s economic consequences (i.e., purely consequential concerns). Hence, we examine whether transmitting false information affects sanctions compared to an analogous report when lying is not involved.

Furthermore, by comparing the effect of dishonesty on punishment across payoff situations, we examine whether punishment for lying remains heterogeneous with respect to the lie’s consequences even after controlling for the sanction attributed to any non-lying preference, which

represent outcomes, the mapping between outcomes and payoff allocations, and the position of the latter, are fully randomized across games and participants.

would be inconsistent with the concept of “pure” lying aversion.

Our preregistered hypotheses, motivated by applying the findings in the lying-aversion literature to the punishment for dishonesty, are:

- H₁) Lying is an aggravating factor for punishment, that is, fixing the economic consequences, the punishment attributed to a report is higher if it is false.
- H₂) The effect of lying on punishment depends on the economic consequences of the lie for the receiver.

Hypothesis H₁ is consistent with lying aversion influencing sanctions for all lie types. Hypothesis H₂ suggests that punishment for dishonesty is not only driven by “pure” lying aversion, (i.e., individuals do not punish dishonesty *per se* only), but some lies are more sanctionable than others beyond their payoff consequences. The combination of Hypotheses H₁ and H₂ indicates that lying has a context-dependent effect on sanctions that is not only driven by consequential concerns, but also by differences in the deontological aspect of dishonesty.

4.2.3 The role of truth-telling in sanctions

This section examines the role of truth-telling in punishment. In the dishonesty literature, truth-telling preferences and lying aversion are often used interchangeably since they cannot be distinguished in settings with a binary message space. Our design allows us to separate the impact of truth-telling on punishment from the effect of lying.

To measure the influence of truth-telling on sanctions, we use the decisions of the Reporting setting when the sender observes the other state. In this scenario, which we refer to as the *Truth Game* (TG), the sender benefits economically from truthfully reporting the observed outcome to the receiver.¹⁵ Hence, the sender earns 15 points for telling the truth and

¹⁵An example of the sender’s decision page in the Truth Game is shown in Figure 69 in Appendix C.9.2.

10 points for lying. The receiver’s payoffs are those in Table 6 with the labels for the message veracity interchanged. Consistent with the notation in the Choosing Game, we refer to the corresponding true messages in this game as harmful, neutral, and beneficial truths. Figure 7 shows an example of the observer’s decision page in the harmful situation.

The **outcome that Person 1 observed** in this situation is: 

The table below shows both participants' payoffs corresponding to each **possible report** Person 1 could make.

	Person 1's payoff	Person 2's payoff
If Person 1 reports 	10	10
If Person 1 reports 	15	5

Please, select below how many **points (0-10)** you would like to deduct from Person 1's earnings for each possible report. These points will be deducted if Person 1 makes this report.

	0	1	2	3	4	5	6	7	8	9	10
If Person 1 reports 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
If Person 1 reports 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Please, press the 'Next' button after completing this page (entering a value for each possible report).

Next

To display the instructions of this task again, click [here](#).

Figure 7: Decision page in the Truth Game (harmful situation) from the observer’s perspective.

Comparing punishment for true reports in the Truth Game with that for their payoff-equivalent messages in the Choosing Game, we study whether truth-telling is a mitigating factor for sanctions. In this regard, our preregistered hypothesis is:

- H₃) Truth-telling mitigates punishment for selfish reports: holding economic consequences fixed, a selfish report is punished less when it results from truth-telling than when it is merely chosen.

Our main interest in truth-telling as a mitigating factor focuses on selfish messages, i.e., those that benefit the sender but generate inequality (either at the expense of the receiver or not). However, truth-telling could also mitigate sanctions for non-selfish choices. For instance, consider a (presumably small) population with an intrinsic preference for punishment (e.g., people who like to damage others) who punish the sender even for a Pareto-beneficial choice. If a fraction of them value truth-telling, its norm-following component could reduce sanctions. Moreover, we can examine if the mitigating role of truth-telling mirrors the aggravating effect of lying on sanctions by comparing both effects across payoff situations.

4.2.4 The Mechanisms Behind Sanctions

Finally, we investigate the determinants of punishment in communication games. In particular, we examine how punishment relates to normative beliefs (both the personal viewpoint and perceived social norms) and to empirical expectations about reporting behavior. In a subsequent part of the experiment, we elicit observers' personal norms (their own assessments of each report's appropriateness), injunctive social norms (their beliefs about what others consider appropriate), and empirical expectations (the expected frequency of senders making each report, also known as descriptive norms).

The elicitation of personal and injunctive norms follows Bašić and Verrina (2024), using a 6-point Likert scale ranging from "very personally appropriate" to "very personally inappropriate" for personal norms, and an analogous scale for injunctive norms using "socially" instead of "personally". Personal and injunctive norms are elicited by asking observers to assess each of the two possible messages. Empirical expectations are elicited by asking observers to indicate (using a slider) the expected number of senders (out of 100) who sent each message. Examples of the decision pages used to elicit personal, injunctive and descriptive norms in the Lying Game are provided in Figures 61, 63, and 65 respectively, presented in Section C.9.1 of the Appendix.

While personal norms are incentivized using a flat rate, injunctive norms are elicited using the coordination method of Krupka and Weber (2013), in which participants receive a bonus (0.5 GBP) if their evaluation matches the modal response of others. Empirical expectations are incentivized with a bonus for guesses within a 5% threshold of the true value. We elicit observers' norms for all games and payoff situations, following the same order used for punishment in the first part of the study.¹⁶

By comparing observers' decisions across the two parts, we examine how personal, injunctive and descriptive norms relate to punishment at the individual level. We also study how lying and truth-telling affect personal and social norms. The main preregistered hypotheses related to these analyses are:

H₄) Personal norms and social norms correlate with punishment.

H₅) The relationship between (dis)honesty and personal and social norms mirrors the relationship between (dis)honesty and punishment. Specifically, lying aggravates normative judgments and empirical expectations: compared to their payoff-equivalent choices, false messages are judged as more inappropriate and are expected to be less frequent. Conversely, compared to their payoff-equivalent selfish choices, selfish true reports are judged as more appropriate and are expected to be more frequent.

Furthermore, we distinguish the role of personal and injunctive norms in punishment, discerning the influence of individuals' own assessments from that expected by the society. In fact, using the payoff situations, we study normative behavior in environments that vary in the number of perceived social norms and their uncertainty.¹⁷ This feature may not only affect normative assessments, but also the relative importance of personal

¹⁶Further details about the different order conditions in the experiment, including the games, payoff situations and different norms, are detailed in Subsection 4.2.5.

¹⁷Recent literature has shown growing interest in normative uncertainty within economic games (D'Adda, Dufwenberg, et al., 2020; Bicchieri, Dimant, Gächter, et al., 2022; Dimant and Gesche, 2023; Dimant, Galeotti, and Villeval, 2024; Dimant, Gelfand, et al., 2025).

versus injunctive norms on sanctions. Our preregistered hypotheses in this regard are:

- H₆) Sanctions for dishonest behavior are more strongly related to what people think should be done (personal and injunctive social norms) than to what they expect others will actually do (descriptive social norms).
- H₇) In situations with a single social norm, there are no significant differences between the influence of personal and injunctive norms on punishment. However, in situations without a unique social norm, sanctions are differently related to personal than to injunctive norms.

In particular, the beneficial situation in the Lying Game provides a setting where not only multiple normative views (either on the personal or the social side) may coexist, but also can point to opposite actions.¹⁸ Accordingly, we examine personal and social norms for both possible actions in this situation, keeping in mind that reporting the truth may be perceived as inappropriate contrary to the predictions of truth-telling preferences or “pure” lying-aversion models.

4.2.5 Experimental procedure

Participants make decisions for all payoff situations within each game. We counterbalance the order of games and situations using a 2×2 between-subject design: one factor determines whether participants first face games with or without lying involved, and the other factor determines which payoff situation (either harmful or beneficial) is presented first (with the rest of situations provided in random order).

Specifically, in one game-order condition (denoted *Reporting first*), participants decide first in all situations of the Reporting setting (LG

¹⁸This differs from the usually studied case where lying has a harmful impact and therefore there is no normative uncertainty since all views point to the norm-following action of telling the truth.

and TG), and then in all situations of the Choosing Game. The order is reversed in the other game-order condition (denoted *Choosing first*). This allows us to control for potential order effects in the differences between lying and payoff-equivalent choices.

In addition, participants first face either a harmful or a beneficial payoff situation in the Lying Game, and the remaining situations are presented in random order. This allows us to compare sanctions for harmful and beneficial lies without any previous experience in the study. Overall, this within-subject design with between-subject order controls measures punishment at the individual level while accounting for potential order effects across games and situations, allowing clean comparisons between harmful and beneficial reports and their payoff-equivalent choices using the first observation in each condition.

To prevent compensation, reputation, or learning effects, payments are determined by the final payoffs in one randomly selected game and situation from the entire study. The receiver earns the payoff corresponding to the report of her randomly matched sender in the selected situation, and the sender earns the corresponding payoff net of the punishment points assigned by the randomly matched observer. Besides, to avoid envy concerns, observers were informed that their earnings in the study were higher than those of the players in the situations they evaluate.

Finally, after making all punishment decisions in the first part of the study (henceforth Part 1), observers conduct the second part, where personal and social norms are elicited. Norms are elicited in blocks: for a given game type, we elicit each norm type—either personal (PN), injunctive (IN) or descriptive (DN)—for all payoff situations before moving to the next norm block. Games and situations follow the same order used in Part 1. To control for order effects across norms types, we fully randomize the blocks' order across participants, having 6 between-subject conditions: PN-IN-DN, PN-DN-IN, IN-PN-DN, IN-DN-PN, DN-PN-IN, DN-IN-PN.

Therefore, each observer is randomly allocated to one of 24 between-subject order conditions ($2 \times 2 \times 6$). Given our main research interests,

the randomization was programmed to ensure a balanced distribution across the main between-subject order conditions: the setting, the first payoff situation, and the norm elicitation blocks. Each observer makes punishment and norm decisions for all games and situations.

Moreover, restricting attention to observers' measures (either punishment or norms) in the first situation they face allows us to treat the data as coming from the corresponding $2 \times 2 \times 6$ between-subject design. Before starting the experiment, each participant was randomly assigned to a single role and order condition. The diagram in Figure 8 graphically summarizes the procedure followed by observers. For the procedure of senders and receivers, see Section C.7 of the Appendix.

In the norms literature, it is standard to elicit norms after the behavior of interest (see Bicchieri, Dimant, and Xiao (2021), Aycinena et al. (2022), Behnk, Hao, and Reuben (2022), and Bicchieri, Dimant, and Sonderegger (2023) for some examples using different economic games). D'Adda, Drouvelis, and Nosenzo (2016) showed that eliciting norms after behavior using a within-subject design does not distort norm measurements, and subsequent studies confirmed this result in other games (Erkut, Nosenzo, and Sefton, 2015; Barr, Lane, and Nosenzo, 2018). Therefore, to get a clean measure of our main outcome of interest (punishment), we elicited observers' norms after they made all punishment decisions.

Furthermore, to measure norms without any influence of sanctions and make a clean between-subject comparison with punishment, we elicit norms in an external sample: participants who later serve as receivers. Specifically, these participants first complete the norm-elicitation tasks in one of the settings (Reporting or Choosing), and then, only in the second part of the experiment, they are assigned the role of receiver for the other setting.

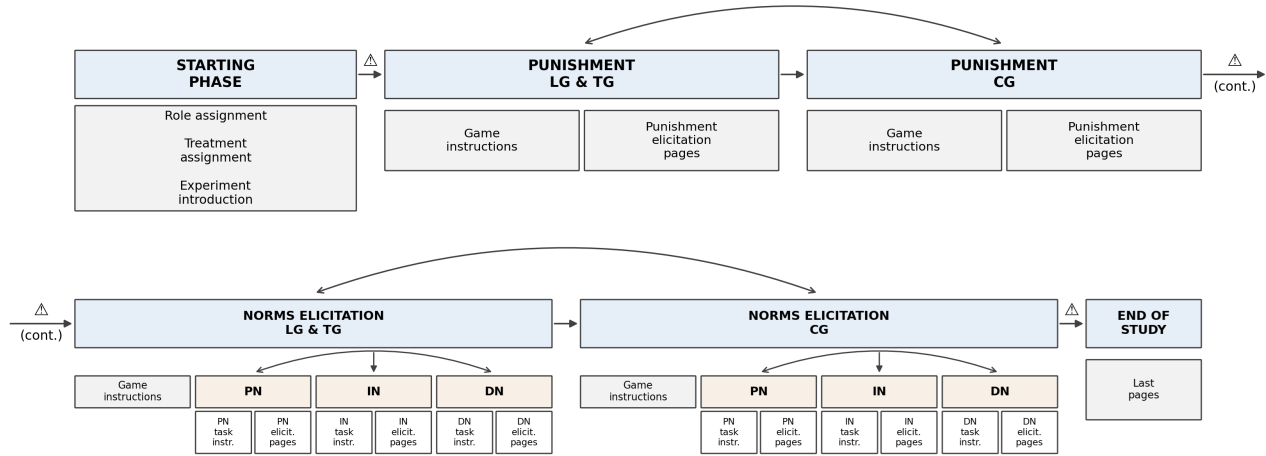


Figure 8: Procedure of observers in the *Reporting first* game-order condition. Curved arrows represent the randomization of the order across conditions. Warning icons indicate where attention checks were conducted.

As a result, half of the receiver sample conducts the norm elicitation for the Reporting setting (i.e., the LG and TG), and the other half does so for the Choosing Game. The rest of the design features, such as the order of situations or norm blocks, are those used to elicit observers' norms.¹⁹ To avoid hedging and income effects, participants were paid based on their earnings in a single part of the experiment. If the norm elicitation part was selected, participants were paid based on the results in one norm block randomly determined.

The experiment was conducted online, and participants were recruited via Prolific (Palan and Schitter, 2018), a platform widely used in economic experiments.²⁰ We recruited a total of 1224 participants (408 assigned to each role), ensuring one-to-one matching and a fully balanced allocation across order conditions. The sample size was determined using a power analysis with an α error probability of 5% and power ($1-\beta$) of 80% to detect small effect sizes (Cohen's $d \approx 0.28$ for the most stringent non-parametric between-subject tests and $d \approx 0.14$ for the most stringent non-parametric within-subject tests).

The sample was composed of participants based in the UK or the US who were fluent in English, and aged between 18 and 50 years (mean age = 32.8).²¹ Within each role, the sample was evenly split by gender (50% men and 50% women).

We included several attention checks during the experiment to ensure data quality and exclude participants who did not read the contents of the study pages. These checks did not assess participants' ability, and were implemented at the beginning and the end of each part of the study. Participants were paid 8 GBP per hour on average, and the experiment

¹⁹For a diagram of receivers' procedure in both conditions, see Figure 51 in Section C.7 of the Appendix.

²⁰Recent studies document high data quality on Prolific relative to other online recruitment platforms (Peer, Brandimarte, et al., 2017; Peer, Rothschild, et al., 2022; Douglas, Ewell, and Brauer, 2023; Albert and Smilek, 2023), and comparable quality to laboratory samples (Gupta, Rigotti, and Wilson, 2021; Suri, Kube, and Schultz, 2025).

²¹All participants had an approval rating (i.e., the percentage of their previous Prolific submissions approved by researchers) of at least 95%. Further information on the sociodemographic characteristics of the participant pool is provided in Section C.8 of the Appendix.

took, on average, around 14.5 minutes to be completed.²²

4.3 Results

Overall, there are no statistically significant differences between the two game-order conditions (Reporting first vs. Choosing first).²³ In particular, punishment in the Reporting setting does not differ significantly depending on whether it is elicited first or second (Mann–Whitney test $p = 0.710$), nor does punishment in the Choosing Game ($p = 0.756$).

This is confirmed by regression analyses that account for individual heterogeneity. Specifically, we estimate linear mixed-effects models with participant-level random effects. There are no order effects when punishment is compared at either the setting level (Wald χ^2 test, $p = 0.280$) or the game level ($p = 0.411$). Regression results are reported in Section C.2 of the Appendix. Therefore, we exploit the within-subject variation in our data. The main conclusions drawn from the results we presented below remain unchanged when using between-subject comparisons, and in some cases are even stronger in the directions discussed. We will specify along the text the few cases where a between-subject comparison leads to a different conclusion.

We begin by describing observers' punishment decisions. Subsections 4.3.1 and 4.3.2 present descriptive statistics and the main non-parametric comparisons.²⁴ Subsection 4.3.3 discerns the influence of lying and truth-

²²This payment rate is in line with the typical rates used in Prolific studies and is considered fair according to Prolific's guidelines. Observers were paid on average 2.55 GBP and they took on median 22 minutes to complete the entire study. Senders were paid on average 1.39 GBP and took on median 8 minutes, whereas receivers were paid on average 1.85 GBP and took on median 15 minutes.

²³When examining game-order effects on punishment separately by payoff situation, we find an order effect only for (Pareto) beneficial lies (Mann–Whitney $p < 0.001$), which are punished more when the Lying Game is presented first. However, since there are no significant differences in punishment for payoff equivalent choices, this effect works against any spurious aggravating effect arising from the within-subject design.

²⁴All tests are two-sided. As preregistered, statistical results shown are corrected for multiple testing using Holm (1979) method. Results are almost identical without correcting

telling on punishment using our preregistered regression framework. Subsection 4.3.4 presents the norms results, and Subsection 4.3.5 analyzes the relationship between norms and punishment. Finally, Subsection 4.3.6 provides sender behavior results.

4.3.1 Punishment for dishonesty

Figure 9 provides, for each payoff situation, a descriptive comparison across games of our main outcome of interest: Punishment for senders' payoff-maximizing messages, i.e., lying in the Lying Game, a payoff-maximizing choice for the sender in the Choosing Game, and truth-telling in the Truth Game.²⁵

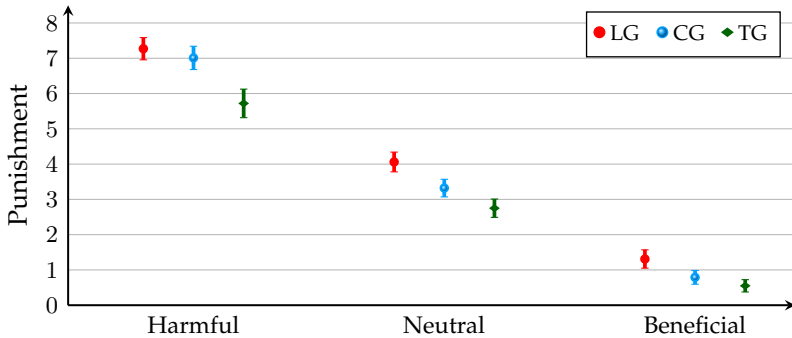


Figure 9: Punishment across games for senders' payoff-maximizing messages (i.e., lying in the LG, truth-telling in the TG and not choosing the 10–10 split in the CG) by payoff situation.

For harmful reports, lying increases sanctions only marginally: punishment for a selfish harmful lie is only marginally different from that for its payoff-equivalent choice (Wilcoxon signed-rank test, $p = 0.06$). Moreover, this small increase (about 3.7%) disappears in the between-subject comparison across game-order conditions (Mann–Whitney test, $p = 0.614$).

for multiple comparisons.

²⁵The results for messages that do not benefit the sender are provided in Section C.1.1 of the Appendix.

By contrast, lying significantly increases sanctions when it is harmless for the receiver: punishment for a false report significantly differs from that for its payoff-equivalent choice for neutral and beneficial lies ($p < 0.001$ in both cases). In particular, lying increases punishment by about 22.2% for neutral reports and by about 66.7% for beneficial reports. These results are maintained when comparing messages using between-subject comparisons.

Moreover, we examine whether dishonesty affects the distribution of punishment. Figure 10 shows the empirical cumulative distribution functions (ECDFs) by game and payoff situation. For all payoff situations, punishment for lying first-order stochastically dominates (henceforth, dominates) that of their payoff-equivalent choices. However, there is no statistically significant difference between the punishment distribution of a harmful lie and a harmful choice (Stuart–Maxwell test, $p = 0.355$).²⁶ Conversely, the punishment distributions for neutral and beneficial lies differ significantly from those for their choices ($p < 0.017$).²⁷

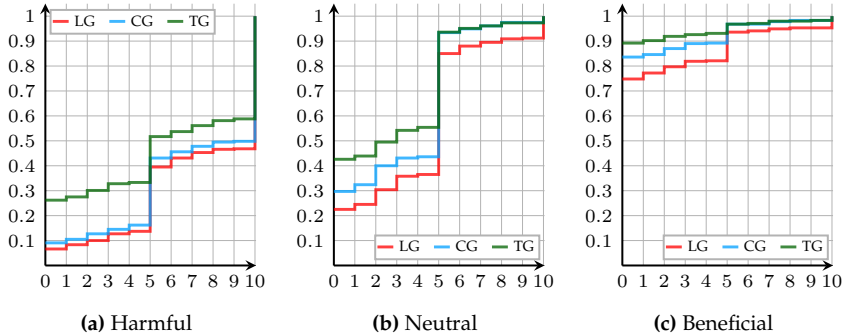


Figure 10: Empirical cumulative distribution functions (ECDFs) of punishment for messages that benefit the sender by game and situation.

²⁶The Stuart–Maxwell test is a generalization of McNemar’s test for marginal homogeneity to outcomes with more than two categories (Sun and Z. Yang, 2008). Our conclusions are preserved using permutation tests with 5000 replications, which account for the data ordinality while exploiting the within-subject variation.

²⁷Using the between-subject comparison, the difference for beneficial lies is only marginally significant (χ^2 test $p = 0.073$).

All in all, these results indicate that the aggravating effect of lying plays a secondary role in punishment, becoming more relevant the fewer outcome-based concerns are violated. Note that while 74.3% of observers punish a false message that benefits the sender at least as much as its equivalent choice, only 3.2% always punish it more. Moreover, the heterogeneity across lie types suggests that the deontological component of dishonesty in punishment varies with the lie's monetary consequences. Therefore, these results confirm Hypothesis H_2 , whereas they partially support Hypothesis H_1 .

Furthermore, consistent with our preregistered secondary hypothesis, punishment for lying decreases with the receiver's payoff, following the expected pattern driven by other-regarding preferences: (selfish) harmful lies are punished the most, whereas (Pareto) beneficial lies are punished the least. The same pattern holds when there is no lying involved in the game. Besides, beneficial reports are rarely punished: 75% (84%) of observers do not punish a beneficial lie (choice). Therefore, our findings indicate that punishment for false reports is primarily driven by the consequences of the lie, rather than by dishonesty *per se*.

Finally, comparing the sanctions of both possible messages in the beneficial situation of the Lying Game, we observe that a beneficial lie is punished significantly more than its truthful (but mutually harmful) alternative (Wilcoxon signed rank test, $p = 0.003$). This suggests that, provided that distributional concerns are not violated, the aggravating effect of dishonesty on punishment may outweigh the impact of inefficiency, pointing to an independent deontological component in punishment decisions.

4.3.2 The influence of truth-telling on sanctions

To investigate the role of truth-telling in sanctions, we compare punishment in the Choosing Game with punishment for payoff-equivalent truthful reports in the Truth Game.

As shown in Figure 9, truth-telling significantly lowers sanctions in all

payoff situations, although not entirely: there are significant differences between the punishment for a true report and that for its payoff-equivalent choice for harmful and neutral truths (Wilcoxon signed-rank test, $p < 0.001$ in both cases). Specifically, truth-telling reduces punishment by about 18.4% for harmful reports and by 17.2% for neutral reports, confirming Hypothesis H₃.

In addition, truth-telling mitigates punishment for beneficial reports: punishment for a beneficial truth is significantly different from that for its payoff-equivalent choice ($p = 0.009$). Since a beneficial choice is already rarely punished (mean = 0.79), this result suggests that truth-telling has a mitigating effect even among participants with an intrinsic preference to punish (e.g., people who like to punish others).

Our hypothesis about the mitigating role of truth-telling is reinforced when comparing punishment distributions. As shown in Figure 10, for all payoff situations, punishment for truthful reports is dominated by that of their payoff-equivalent choices. These distributional differences are statistically significant for harmful and neutral truths (Stuart–Maxwell test, $p < 0.001$ in both cases), and for a beneficial truth ($p = 0.039$), although the latter becomes statistically not significant with a between-subject comparison ($p = 0.113$).

However, truth-telling does not eliminate punishment entirely. Even if it nearly does so for beneficial reports, truth-telling is still substantially punished when it creates inequalities in favor of the sender. This is particularly evident for a selfish harmful truth: only about 26% of observers refrain from punishing it, whereas 41% punish it to the full extent.

Furthermore, punishment in the Truth Game is concentrated at levels that implement equal payoffs for the two players, reinforcing the interpretation that distributional considerations are crucial for punishment decisions. Together with the results from the dishonesty section, and with the fact that reports that do not benefit the sender tend to be sanctioned significantly less (see Figure 39 in Section C.1.1 of the Appendix), our findings suggest that distributional concerns are the main consequential factor

behind punishment, particularly those related to preferences for equality. This interpretation is consistent with previous evidence from studies on other games, where models that incorporate inequity-aversion concerns best explain the occurrence and magnitude of sanctions (Leibbrandt and López-Pérez, 2012; Grundherr, Jauernig, and Uhl, 2021).

4.3.3 Discerning lying aversion from truth-telling preferences

Overall, our results indicate that punishment is primarily driven by the report’s consequences. Truth-telling substantially mitigates punishment, whereas lying plays a secondary role that becomes more relevant in the absence of harmful consequences. We now quantify the effects of lying and truth-telling on sanctions using our regression analyses.²⁸

To examine the influence of lying on punishment overall and by situation, and quantify the mitigating role of truth-telling, we estimate the following linear model at the report level, controlling for heterogeneity at the individual level using fixed effects.

$$PUN_{ijgs} = \alpha + \beta_1 S_{10j} + \beta_2 LG_g + \beta_3 TG_g + \beta_4 (S_{10j} \times LG_g) + \beta_5 (S_{10j} \times TG_g) + \mu_i + \epsilon_{ijgs} \quad (4.1)$$

where PUN_{ijgs} is the punishment assigned by observer i in game g and situation s if the sender reports message j . S_{10} is a dummy variable equal to 1 when the sender does not report their payoff-maximizing message. LG_g is a dummy equal to 1 if the game corresponds to the Lying Game, TG_g is the analogous indicator for the Truth Game, and the Choosing Game is the baseline category. μ_i denotes observer fixed effects. We estimate this model with a regression for each payoff situation and pooling across situations all observations.

Therefore, β_2 identifies the net effect of lying on punishment for the

²⁸Unlike the non-parametric tests in Sections 4.3.1 and 4.3.2, these analyses use the pooled punishment data.

sender's payoff-maximizing reports, that is, after controlling for the lie's economic consequences and any preferences unrelated to lying. Likewise, β_3 captures the net effect of truth-telling on punishment for the sender's payoff-maximizing reports. By contrast, β_4 identifies the net effect of truth-telling for reports that do not benefit the sender, whereas β_5 captures the net effect of lying for reports that do not benefit the sender. Table 7 reports the estimates from Model (4.1) both separately by payoff situation and for the pooled sample across situations.

D.V: Punishment	Harmful	Neutral	Beneficial	All
LG	0.260* (0.138)	0.738*** (0.132)	0.525*** (0.132)	0.507*** (0.090)
TG	-1.297*** (0.195)	-0.569*** (0.134)	-0.240** (0.095)	-0.702*** (0.099)
S_{10}	-6.596*** (0.190)	-2.745*** (0.149)	0.257** (0.125)	-3.028*** (0.108)
$LG \times S_{10}$	-0.240 (0.162)	-0.824*** (0.148)	-0.723*** (0.164)	-0.596*** (0.111)
$TG \times S_{10}$	1.873*** (0.249)	1.108*** (0.190)	0.922*** (0.161)	1.301*** (0.144)
Mean punishment baseline	7.01	3.32	0.79	3.71
Number of observations	2448	2448	2448	7344
Adj. R^2	0.621	0.340	0.042	0.210

Notes: Fixed-effects regression results with clustered standard errors at the participant level (in parentheses). Baseline category is CG when $S_{10} = 0$, i.e., choices in the CG that benefit the sender. LG and TG denote that the punishment is in the Lying Game and Truth Game. S_{10} indicates that the sender does not benefit from the message sent.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 7: Punishment regression results by situation and the pooled sample.

Pooling all situations, we find a significant aggravating effect of lying on punishment on average ($p < 0.001$). Specifically, lying increases the punishment for messages that benefit the sender by about half a point relative to the baseline choice. However, when discerning the effect for each situation separately, we find that the magnitude of lying's influence is small and marginally significant for a harmful lie (0.26, $p = 0.059$).

In contrast, lying significantly increases punishment for neutral (0.74, $p < 0.001$) and beneficial (0.53, $p < 0.001$) reports. In fact, when looking at the effects of lying on punishment for reports that do not benefit the sender (i.e., the estimate of β_5), we observe that dishonesty increases punishment either when the receiver is not affected (1.11, $p < 0.001$), or when the receiver is better-off economically (1.87, $p < 0.001$). Overall, these results suggest that the role of lying in sanctions becomes stronger as the message's consequences become more favorable to the receiver.

However, inequity favoring the sender has the strongest influence on sanctions: on average, punishment increases by 6.596 points when the sender benefits at the receiver's expense, and by 2.745 points when the receiver is economically unaffected.

Furthermore, we estimate the differences of lying's influence on punishment across lie types.²⁹ On average, the aggravating effect of lying is significantly smaller for harmful than for neutral reports ($p < 0.001$). However, there is no significant difference between beneficial and neutral ($p = 0.177$) or harmful ($p = 0.142$) lies. These results suggest that the increment in punishment driven by lying *per se* depends on the monetary consequences that a lie entails.

On the other hand, there are asymmetries in the effect of truth-telling on punishment compared to the aggravating role of lying. On average, there is an overall aggregate mitigating effect of truth-telling on punishment (-0.702, $p < 0.001$). This mitigating effect is present in all situations: for harmful reports, truth-telling has a big lowering effect (-1.297, $p < 0.001$). For neutral reports, its effect is still large (-0.569, $p < 0.001$), whereas it is smaller in magnitude for beneficial reports (-0.240) but still statistically significant ($p = 0.011$).³⁰

²⁹To estimate differences in punishment across payoff situations, we include situation indicators and their interactions. The expanded specification and its regression results are reported in Section C.3 of the Appendix.

³⁰Since the regressions in Table 7 control for individual heterogeneity via fixed effects, order effects that are constant within subjects are ruled out. However, since the order of the settings may affect punishment for each game differently, we extend the regression specification controlling for game-order effects. All our main results are preserved. Results

Finally, we estimate the differences of truth-telling mitigation across situations. The effect of truth-telling significantly differs across situations, following the expected pattern according to our hypotheses: On average, truth-telling has a stronger mitigating role for selfish choices (being the strongest for those that harm the receiver), whereas its influence is weakest for choices that benefit the receiver.

Overall, the results from our regression framework confirm the findings from the previous non-parametric analyses. Our evidence suggests that the deontological norm-breaking aspect of dishonesty is not the main driver of punishment, as it is masked by the stronger effect of outcome-based norm violations. Lying plays a secondary role, becoming more salient as the consequences for the receiver become more favorable. Conversely, truth-telling has a substantial mitigating influence on punishment, although it does not eliminate sanctions for selfish messages. This suggests that the norm-following aspect of truth-telling attenuates the negative effect of selfish reports arising from consequential concerns.

Taken together, these findings show that the message's economic consequences are the main determinant of punishment in communication games. In particular, sanctions are primarily driven by whether the message benefits the sender relative to the receiver, thereby generating inequality, and by whether it makes the receiver worse off.

4.3.4 Norms and Expectations about (Dis)Honesty

In this section, we study the role of lying and truth-telling in personal, injunctive and descriptive norms.³¹ In line with Hypothesis H₅, the effects of lying and truth-telling on personal and social norms tend to mirror the effects on punishment. Figure 11 presents personal norms, injunctive norms and empirical expectations (descriptive norms) for senders'

of the extended regressions are explained in Subsection C.2.1 of the Appendix.

³¹As it is standard in the norms literature, personal and injunctive norms are coded on a 6-point scale from -1 (very inappropriate) to 1 (very appropriate). Descriptive norms are elicited as integers from 0 to 100, with the two percentages within each situation summing to 100.

payoff-maximizing messages. The corresponding figures of personal and injunctive norms for the remaining messages are provided in Section C.1 of the Appendix.

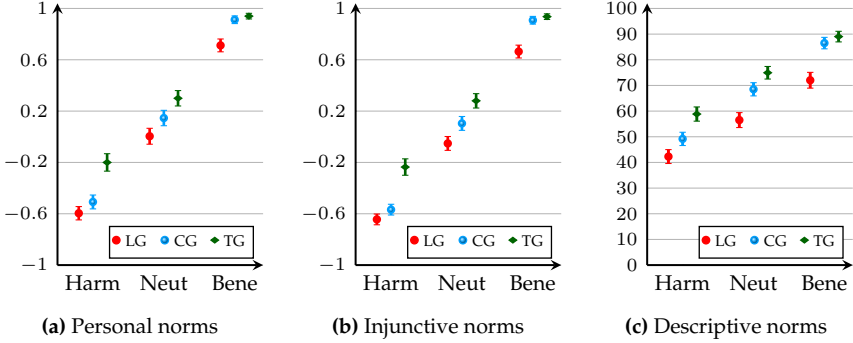


Figure 11: Personal, injunctive, and descriptive norms of sender-benefiting messages. Descriptive norms are the expected number of senders (out of 100) who make the report.

Consistent with the influence of dishonesty on punishment, false reports are judged to be more inappropriate and are expected to be less frequent than their payoff-equivalent choices (Wilcoxon signed-rank test $p < 0.001$ for all comparisons). The only exception relative to punishment concerns harmful lies, which are significantly (rather than only marginally) more inappropriate (both personally and socially) than their payoff-equivalent choices, and they are expected to be less frequent.

In contrast, consistent with the mitigating role of truth-telling in punishment, truthful reports are judged more appropriate and are expected to be more frequent than their payoff-equivalent choices ($p < 0.001$ for all comparisons). The only difference relative to punishment is that beneficial truth-telling has only a marginal effect on norms ($p < 0.062$).

All findings are confirmed when comparing the full distributions of norms (Stuart–Maxwell $p < 0.004$ for all comparisons except beneficial truths, where $p > 0.407$).³² Results are even closer to the effects on

³²Figures reporting the ECDFs of each norm for all games and payoff situations are

punishment under between-subject comparisons: Lying worsens all norm measures when it is beneficial (Mann–Whitney $p < 0.001$), but there is no effect neither on assessments ($p > 0.7$) nor on empirical expectations ($p = 0.231$) for harmful lies. By contrast, sender’s payoff maximizing reports are expected to be more frequent and judged more appropriate when they are true only if the receiver is not benefited ($p < 0.001$).

Moreover, we estimate the influence of lying and truth-telling in personal and social norms using regression analyses analogous to Model (4.1) with each norm type instead of punishment as the dependent variable.³³

The aggravating effect of lying on personal and injunctive norms is statistically significant in all payoff situations. However, its magnitude is modest: on average, assessments do not change direction nor intensity. The largest effect arises for beneficial reports, where lying lowers personal (injunctive) norms by 0.20 (0.24) on average. However, since beneficial choices are judged very appropriate (mean personal norm = 0.912, mean injunctive norm = 0.907), the lying decrease does not overturn the overall assessment. In fact, only 11% (12.5%) of observers consider a beneficial lie as somewhat inappropriate or worse from a personal (social) perspective.

Regarding empirical expectations, on average, reports that benefit the sender are expected to be less frequent when they are false. Across situations, lying reduces the expected frequency of senders reporting the payoff-maximizing message by 14% to 17.5%. However, expectations for beneficial lies remain high: around 79% of observers expect more than half of senders to tell a beneficial lie, and 50% expect at least 90% to do so. Conversely, 37% (58%) of observers expect more than half of senders to tell a harmful (neutral) lie.

By contrast, truth-telling increases personal and social norms when the receiver does not benefit from the report. However, this effect is prominent only for harmful choices, where it attenuates its inappropriateness (0.31

provided in Section C.1.2 of the Appendix, where differences in distributions are discussed in more detail.

³³The model specifications and estimation results are provided in Section C.4 of the Appendix.

for personal norms and 0.33 for injunctive), although on average it does not make it appropriate (neither personally nor socially). In fact, 62.3% (64.7%) of observers consider a harmful truth as somewhat inappropriate or worse from a personal (social) perspective.

Consistent with punishment results, inequality in favor of the sender is the strongest determinant of personal and social norms, affecting evaluations not only in magnitude but also in direction. In the harmful situation, when the sender benefits at the receiver's expense, average assessments shift from appropriate to inappropriate, regardless of whether the message is truthful. When the receiver's payoff is unaffected by the sender's report, payoff-maximizing messages are judged less appropriate, but still remain on average somewhat appropriate.

Furthermore, although a Pareto-beneficial lie is evaluated less favorably than its equivalent non-lying messages, it is still considered appropriate by most observers. At the same time, the mutually harmful alternatives in the beneficial situation are also judged appropriate by a majority of observers. Although this latter finding may seem surprising at first glance, it further supports our interpretation that observers take into account not only whether a report is harmful in absolute terms, but also how it affects the distribution of payoffs between the sender and the receiver. Even though these mutually harmful reports leave the receiver worse off than their Pareto-beneficial alternative, they reduce the sender's payoff by the same amount. As a result, many observers do not regard them as inappropriate, suggesting that distributional concerns are the main driving factor behind norms.

We also examine whether appropriateness judgments exhibit heterogeneity in each situation, or instead reflect a clear consensus around a single norm. Figures 12 and 13 show the histograms of personal and injunctive norms for the messages that benefit the sender in each game and situation.³⁴

³⁴The corresponding figures for the alternative messages are presented in Subsection C.1.1 of the Appendix.

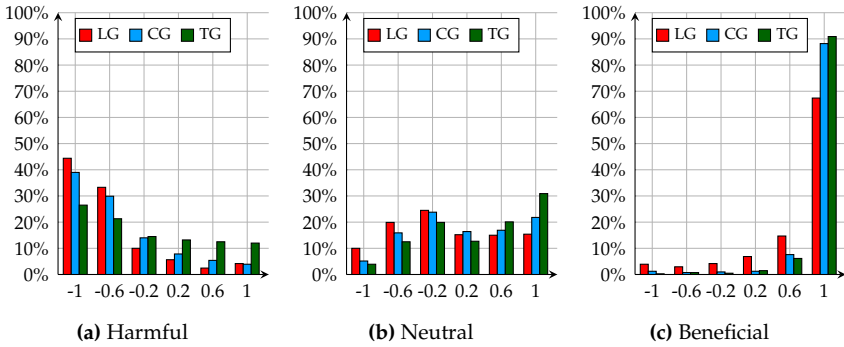


Figure 12: Distribution of personal norms of sender-benefiting messages.

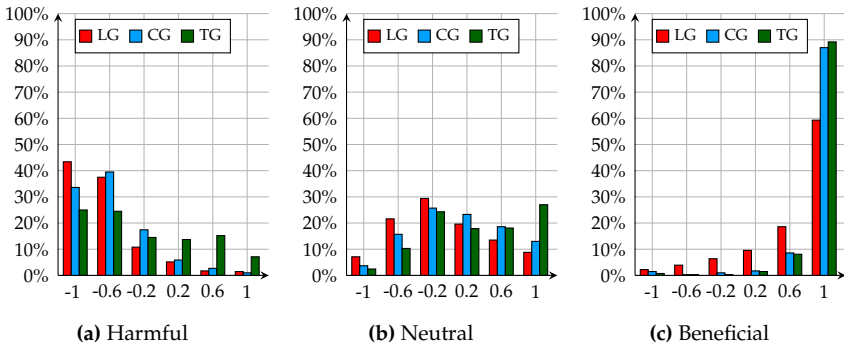


Figure 13: Distribution of injunctive norms of sender-benefiting messages.

On average, beneficial lying is judged as appropriate, both from a personal and a social perspective. Although both norms are lower than those for the choosing and truth counterparts (which are very appropriate for more than 87% of observers), a beneficial lie is still considered personally (socially) very appropriate by 67% (59%) of observers. Conversely, only a small fraction of observers (around 20%) consider beneficial lying inappropriate, whereas this share is negligible when choosing or telling the truth.

On the other hand, there is no clear consensus for harmful truths: A

small majority of observers ($\sim 60\%$) consider it inappropriate, although judgments are widespread. Hence, there is no consensus on a single appropriate rating, and multiple norms arise: While a fraction of observers give more weight to the inappropriate consequences of the message, another significant fraction judge based on the appropriateness of telling the truth. Conversely, most participants consider inappropriate (or worse) the choosing and false counterparts.

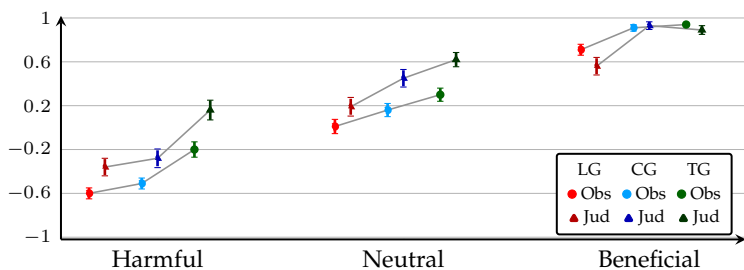
Finally, personal and injunctive norms have no consensus for neutral reports. While truthfulness and lying shift judgments to appropriateness and inappropriateness correspondingly, there is no salient norm arising. We can consider that if any, multiple norms for both being appropriate and inappropriate are considered, although they have limited salience.

Beyond the within-subject evidence on norms and dishonesty explained above, we now examine the robustness of our main conclusions comparing norms in the observer sample with norms elicited from participants who did not make punishment decisions before, who we denote judges for comparability. Figure 14 summarizes these comparisons.³⁵

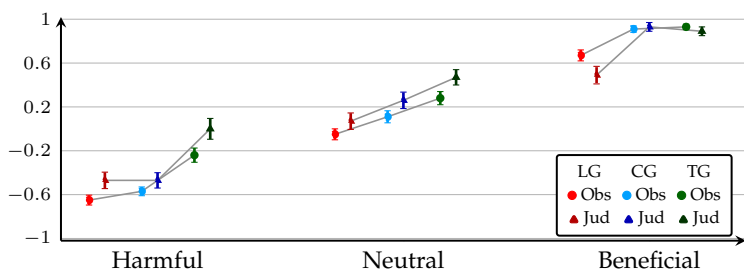
Compared to observers, judges consider harmful and neutral messages that benefit the sender more inappropriate. This difference is statistically significant in all games. Conversely, judges consider beneficial lies more appropriate than observers. For descriptive norms, there are no statistically significant differences between the two samples. These results are robust to using between-subject comparisons only with the observations from the first setting played by observers, which are provided in Subsection C.6.1 of the Appendix.

Although the punishment task may affect normative assessments, the differences in norms driven by dishonesty are preserved when the between-subject comparisons are replicated in the judges' sample. The only exception, relative to the results for observers, concerns harmful lies, which are not evaluated as significantly different from their payoff-

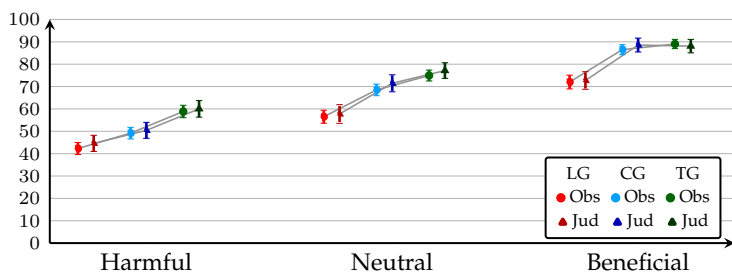
³⁵For readability, the corresponding descriptive tables and tests comparing the two samples are presented in Section C.6.



(a) Personal norms.



(b) Injunctive norms.



(c) Descriptive norms.

Figure 14: Comparison of norms with and without previous punishment elicitation using all observations. Means and 95% confidence intervals are shown for observers (with punishment) and judges (without punishment) across situations and games. The top, middle, and bottom panels display personal (PN), injunctive (IN), and descriptive norms (DN).

equivalent choices ($p = 0.16$ for personal norms, and $p = 0.819$ for injunctive norms). Besides, beneficial truth-telling has no significant effect in the distribution of personal ($p = 0.219$) or injunctive norms ($p = 0.195$).

Therefore, the effects of lying and truth-telling on punishment align more closely with their effects on norms in the judges' sample than in the observers' sample. This reinforces the credibility of our findings on lying and truth-telling influencing norms, indicating that the elicited norms do not merely justify previous punishment decisions.

4.3.5 The role of norms in punishment

Finally, we investigate how personal and social norms relate to punishment. First, we provide a between-subject comparison of judges' norms and observers' punishment. Specifically, we compute the averages of judges' norms and observers' punishment by game, situation, and the message's economic consequences. Table 8a provides the correlation matrix based on these 18 observations. Average punishment is not significantly correlated with average descriptive norms (0.107, Pearson product-moment correlation test $p = 0.67$). By contrast, average punishment is strongly negatively correlated with average personal norms (-0.760) and average injunctive norms (-0.878), with both correlations statistically significant ($p < 0.001$).

We find a similar pattern when using observers' average norms instead (see Table 8b): the correlations between average punishment and average personal and injunctive norms are stronger, whereas the correlation with average descriptive norms is weaker.

Using the full observer sample (i.e., the individual-level observations instead of observers' averages), the same pattern emerges, although the correlations are weaker. Table 8c provides the correlation matrix based on all individual-level observations rather than on message-level averages. Punishment remains negatively correlated with personal (-0.58) and injunctive (-0.58) norms (Spearman correlation test $p < 0.001$ in both cases),

	PUN	PN	IN	DN		PUN	PN	IN	DN		PUN	PN	IN	DN
PUN	1				PUN	1				PUN	1			
PN	-0.76	1			PN	-0.93	1			PN	-0.58	1		
IN	-0.88	0.97	1		IN	-0.94	0.99	1		IN	-0.58	0.79	1	
DN	0.11	0.51	0.30	1	DN	-0.02	0.27	0.25	1	DN	-0.04	0.26	0.21	1
(a) With cell means of observers' punishment and judges' norms					(b) Using only observers' cell means					(c) Using all observers data				

Table 8: Correlation matrices for punishment (PUN), personal norms (PN), injunctive norms (IN), and descriptive norms (DN) under alternative samples and levels of aggregation. Panel (a) combines cell-level average punishment from observers with cell-level average norms from judges, where each cell is defined by the combination of (game, situation, message). Panel (b) uses the 18 cell-level means computed within the observer sample. Panel (c) uses all observations of observers instead of the cell-means.

whereas its correlation with descriptive norms is negligible (-0.05).

Moreover, consistent with findings in previous work, personal and injunctive norms are strongly correlated (0.78 in the Lying Game, 0.81 in the Choosing Game, 0.78 in the Truth Game, and 0.79 in the pooled sample, $p < 0.001$). By contrast, both appropriateness norms are much less correlated with empirical expectations (always below 0.31, $p < 0.001$).

Therefore, we study the relation between punishment and personal and social norms with the following model, using all observers' within-subject decisions.³⁶ Following previous work on the influence of norms in behavior, we model punishment as linearly related to personal and social norms (Capraro and Perc, 2021; Bašić and Verrina, 2024).³⁷

³⁶Since two of the main predictors (personal and injunctive norms) are strongly correlated, we assess potential multicollinearity issues using variance inflation factors (VIFs). Across all regression models, all VIFs are below 5, which indicates that concerns of multicollinearity are unlikely driving our results.

³⁷To assess the individual predictive power of each norm, we also estimate reduced specifications including each norm separately (PN-only, IN-only, DN-only), and in pairs. These results are presented in Section C.5 of the Appendix. While using both personal and injunctive norms increases the predicted power compared to any single-norm model,

$$PUN_{ijgs} = \alpha + \theta_{PN} PN_{ijgs} + \theta_{IN} IN_{ijgs} + \theta_{DN} DN_{ijgs} + \mu_i + \epsilon_{ijgs} \quad (4.2)$$

Table 9 provides the estimates by situation.³⁸ Across all payoff situations, personal and injunctive norms are strongly associated with punishment, whereas descriptive norms have practically no relation, with small coefficients even when statistically significant.

D.V: Punishment	Harmful	Neutral	Beneficial	All
PN	0.966*** (0.092)	0.552*** (0.067)	0.265*** (0.074)	0.686*** (0.058)
IN	0.687*** (0.092)	0.537*** (0.066)	0.205*** (0.069)	0.669*** (0.056)
DN	0.004 (0.003)	0.011*** (0.002)	0.000 (0.001)	0.010*** (0.001)
Number of observations	2448	2448	2448	7344
Adj. R ²	0.611	0.377	0.084	0.449

Notes: Fixed-effects regression results with clustered standard errors at the participant level (in parentheses). Personal and injunctive norms coefficients represent the effect on punishment of a one-category increase in perceived inappropriateness. Descriptive norms coefficients represent the effect on punishment of a one percentage point increase in the expected frequency of the message. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 9: Fixed-effects regressions of punishment on personal, injunctive, and descriptive norms.

Comparing the effect of each norm, both appropriateness norms are more strongly associated with punishment than empirical expectations (for all $p < 0.009$). Hence, sanctions are more closely related to what people think should be done (personal and injunctive norms) than to what they expect others will do (descriptive norms), consistent with Hypothesis H₆.

Model (4.2) performs best.

³⁸For the regression analyses in this section, personal and injunctive norms are rescaled from 0 (very appropriate) to 5 (very inappropriate), so that coefficients can be interpreted as the change in punishment associated with a one-category increase in perceived inappropriateness (personal or social).

Nevertheless, both appropriateness assessments have a similar influence ($p > 0.1$ in all cases), although personal norms are directionally more strongly associated with punishment in the harmful situation ($p = 0.12$). However, the relationship between punishment and norms may reflect differences driven by the veracity of the message and its consequences, as the normative implications of a report vary with these factors. This may be particularly relevant when appropriateness norms are heterogeneous, because the presence of multiple injunctive norms may weaken the link between perceived social appropriateness and punishment, relative to the individuals' own assessment.

Therefore, using Model (4.3), we estimate a joint specification to examine the influence of lying and truth-telling on punishment once we account for normative assessments and expectations of the message.

$$\begin{aligned}
 PUN_{ijgs} = & \alpha + \beta_1 S_{10_j} + \beta_2 LG_g + \beta_3 TG_g \\
 & + \beta_4 (S_{10_j} \times LG_g) + \beta_5 (S_{10_j} \times TG_g) \\
 & + \beta_{PN} PN_{ijgs} + \beta_{IN} IN_{ijgs} + \beta_{DN} DN_{ijgs} \\
 & + \mu_i + \epsilon_{ijgs}.
 \end{aligned} \tag{4.3}$$

Table 10 provides the estimation results. Comparing these results with those in Tables 7 and 9 suggests that normative assessments (particularly personal norms) account for a substantial share of the influence of lying and truth-telling on punishment: Once norms are included, the coefficients on lying, truth-telling, and the message consequences become remarkably smaller. In particular, lying is no longer statistically significant overall, whereas the effect of truth-telling is substantially reduced but remains significant. However, although the effect of economic consequences is also attenuated, they remain the main driver of sanctions.

These results indicate that norms explain a substantial part of the mitigating effect of truth-telling on punishment, although not all of it, whereas the aggravating effect of lying mostly operates through differences in the normative assessment of the message. By contrast, although the influence

D.V: Punishment	Harmful	Neutral	Beneficial	All
LG	0.066 (0.144)	0.443*** (0.136)	0.267** (0.129)	0.052 (0.095)
TG	-0.653*** (0.181)	-0.294** (0.134)	-0.205** (0.098)	-0.217** (0.098)
S ₁₀	-4.001*** (0.346)	-1.900*** (0.211)	-0.413** (0.164)	-1.803*** (0.128)
LG × S ₁₀	0.069 (0.175)	-0.362** (0.165)	-0.213 (0.164)	0.206* (0.117)
TG × S ₁₀	0.783*** (0.238)	0.423** (0.181)	0.690*** (0.163)	0.217 (0.154)
PN	0.665*** (0.085)	0.438*** (0.069)	0.264*** (0.074)	0.609*** (0.056)
IN	0.101 (0.091)	0.221*** (0.068)	0.159** (0.070)	0.489*** (0.052)
DN	-0.005* (0.003)	-0.003* (0.002)	-0.002 (0.002)	-0.004*** (0.001)
Number of observations	2448	2448	2448	7344
Adj. R ²	0.668	0.449	0.097	0.487

Notes: Fixed-effects regression results with clustered standard errors at the participant level (in parentheses). Baseline category is CG for messages that benefit the sender.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 10: Regressions of punishment on game, message type, and norms.

of the message's consequences is attenuated once we control for its appropriateness, it remains strong and continues to be the main determinant of punishment. This suggests that the effect of the message's consequences on punishment does not operate solely through their influence on the perceived appropriateness of the message, but also reflects the role of other factors, such as other-regarding preferences.

On the other hand, although the relationship between punishment and norms weakens once we control for the influence of communication and the message's economic consequences, norms remain significantly related to punishment. In particular, personal norms remain robust, whereas

injunctive norms fall much more markedly, becoming statistically insignificant in the Harmful situation. This suggests that injunctive norms mostly capture variation in punishment driven by lying, truth-telling and the message's economic consequences, whereas personal norms remain associated with punishment beyond these factors. As in the previous results without the controls, descriptive norms are mostly unrelated to punishment, and if there is any relation, its magnitude is very small.

Finally, differences in how personal and injunctive norms relate to punishment depend on the degree of consensus about the message's appropriateness. When there is broad agreement about whether a message is appropriate, personal and injunctive norms have a similar relationship with punishment. This is the case in the beneficial situation, where the Pareto-beneficial message is mostly viewed as appropriate, and both norms are similarly related to punishment ($p = 0.412$). By contrast, personal norms are more strongly related to punishment in situations where there is no clear consensus about the message's appropriateness. This is the case for harmful truths ($p < 0.001$) and for messages that benefit the sender without affecting the receiver, although in the latter case the difference is only marginally significant ($p = 0.07$).

One possible interpretation is that, when there is consensus, personal and injunctive norms are closely aligned. By contrast, when different normative views coexist, participants are less likely to adhere to a social norm because they do not perceive a clear injunctive standard. As a result, observers are more likely to rely on their own evaluation of the message, and punishment becomes more closely related to personal norms.

4.3.6 Senders' behavior

Consistent with previous findings in the dishonesty literature, we observe lying aversion across all types of lies. Figure 15 displays the fraction of senders' payoff-maximizing messages by game and situation.

Senders are significantly less likely to report a payoff-maximizing

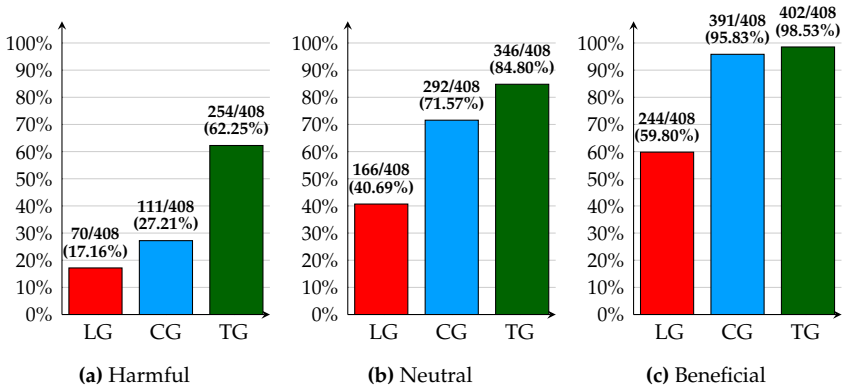


Figure 15: Frequency of senders' payoff-maximizing reports by game and situation.

message when doing so requires lying (McNemar's test $p < 0.001$ in all cases). In particular, dishonesty decreases the probability of such messages from 10 to 36 percentage points, depending on the consequences for the receiver.³⁹

On the other hand, senders are significantly more likely to send a selfish message when it is truthful ($p < 0.001$). However, truth-telling is not enough when it involves sending a selfish message: 37.75% of senders lie to avoid harming the receiver, whereas 15.2% do so to avoid generating a selfish inequality. One interpretation for this behavior is that senders anticipate observers' punishment to be driven primarily by the message's consequences rather than by dishonesty *per se*.

Furthermore, using messages across all games within a given payoff situation, we distinguish lying averse senders (i.e., those who avoid a selfish report if it would be dishonest) from senders who exploit a truth-telling opportunity to send their payoff-maximizing message (suggesting that truth-telling motives outweigh other-regarding concerns). Figure 16

³⁹Since reporting rates in the Choosing Game differ across situations, relative reductions are more informative for cross-situation comparisons than changes in percentage points. The effect of lying is broadly similar across situations, with relative reductions of about 37% for harmful and beneficial messages and 43% in the neutral situation.

displays senders' messages across the three games for each payoff situation, linking each individual's messages from one game to the next.

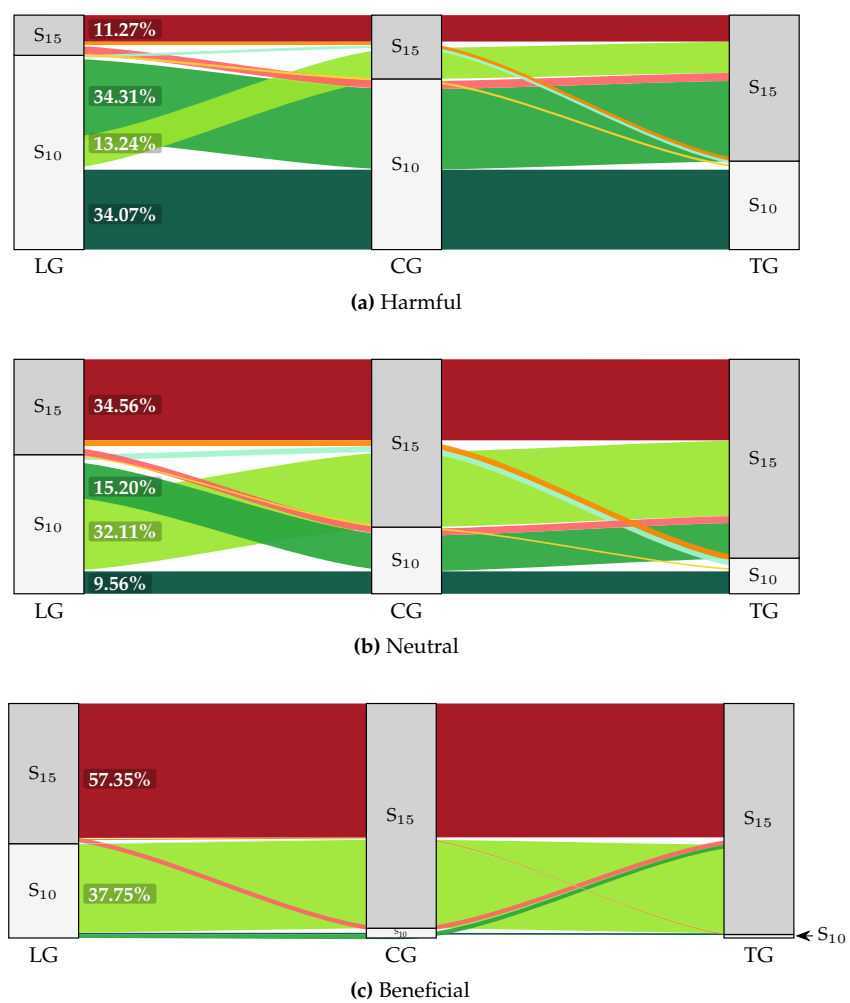


Figure 16: Alluvial plots of senders' reports across games in each situation. S_{15} denotes that the report is payoff-maximizing for the sender, whereas S_{10} denotes that it is not. Flows are labeled only when their frequency is larger than 9%. Labels indicate the fraction of senders who make the corresponding decisions in the three games in the entire sample.

We observe that the fraction of lying-averse senders (those who avoid sending the payoff-maximizing report only when it would require lying) increases as the message becomes more favorable to the receiver, rising from 13.24% when the report is harmful to 37.75% when it is beneficial. This indicates that lying aversion is itself context-dependent.

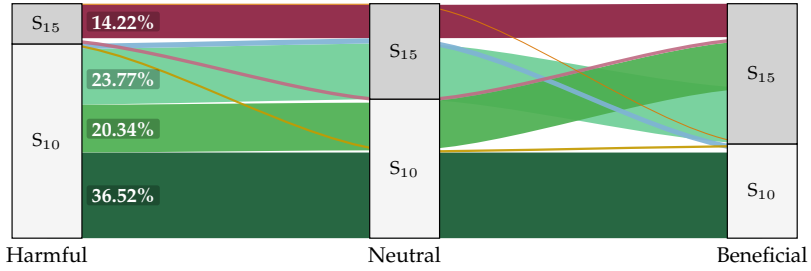
The opposite pattern emerges for truth-telling motives. In particular, 34.31% of senders exploit a truth-telling opportunity to make a selfish harmful report, suggesting that truth-telling serves as a justification for selfish choices. These shifts are also reflected in stable reporting patterns across games. The fraction of senders who always send the payoff-maximizing message rises from 11.27% when it harms the receiver to 57.35% when it is beneficial, whereas 34.07% of senders always choose the payoff-equating message in the harmful situation, rather than benefiting at the receiver's expense. Altogether, these four behavioral patterns account for most sender choices (more than 91%).

Finally, we identify the fraction of "pure" lying aversion in the senders sample, and distinguish senders who are lying averse depending on the lie's economic consequences. Figure 17 displays reporting behavior across the three payoff situations within each game.

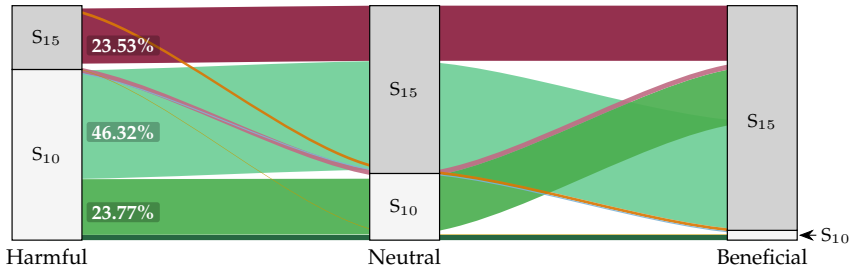
In the Lying Game, 36.52% of senders never lie regardless of the report's consequences, whereas 44.11% exhibit partial lying aversion: 23.77% avoid lying only when it is harmful, whereas 20.34% do not lie when it creates an inequality. By contrast, 14.22% of senders always send their payoff-maximizing false report, whereas 59.31% do so when this report is truthful.

Moreover, 25% of senders forgo telling a self-serving truth when it harms the receiver, whereas 11.27% do so if it creates inequality in their favor. In addition, behavior in the Choosing Game is broadly consistent with other-regarding preferences: 46.32% of senders choose their payoff-maximizing message only when it is harmless, whereas 23.77% always choose the egalitarian message.

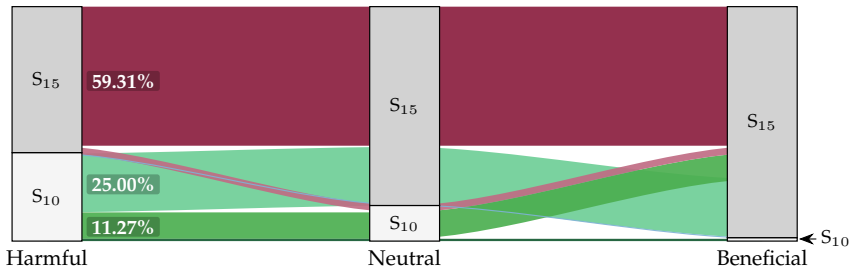
Lastly, considering the reports of all games altogether, 30.88% of



(a) Lying Game (LG)



(b) Choosing Game (CG)



(c) Truth Game (TG)

Figure 17: Alluvial plots of reports across situations for each game. S_{15} denotes that the report is payoff-maximizing for the sender, whereas S_{10} denotes that it is not. Flows are labeled only when their frequency is larger than 9%. Labels indicate the fraction of senders who make the corresponding decisions across the three situations in the entire sample.

senders never send a false report in the entire experiment, consistent with “pure” lying aversion and truth-telling preferences. On the other hand, 27.21% of senders always send the message that is beneficial for the receiver, consistent with other-regarding preferences, whereas only 9.56% of senders always make their payoff-maximizing report. These results reinforce the evidence in previous studies on the role of lying aversion and other-regarding preferences in communication, showing that while both play a substantial role, they complement each other.

4.4 Conclusion

Punishment serves as a regulatory mechanism and has been shown to act as a norm enforcer in cooperative and strategic settings. However, in information-transmission games, punishment has mainly been studied for its role in promoting truthfulness, whereas little is known about what drives such punishment. This paper studies how dishonesty, truth-telling and the report’s economic consequences shape punishment. Using a within-subject experiment, we disentangle the role of dishonesty from consequential factors and other motives unrelated to lying *per se*.

Our results show that, in contrast to our evidence on lying aversion from the sender’s side, punishment for dishonest communication is primarily driven by the message’s economic consequences. The deontological component of lying plays a secondary role, becoming more salient when consequential concerns are weaker. Even conditional on the message’s consequences, the effect of lying on punishment is heterogeneous across lie types, where 75% of participants do not punish a Pareto-beneficial lie. By contrast, truth-telling mitigates punishment, especially when the receiver does not benefit economically. However, it does not eliminate sanctions entirely since its influence does not overcome the punishment given by distributional concerns.

Similar patterns arise for normative assessments: while lying aggravates and truth-telling mitigates the message’s personal and social ap-

appropriateness, the message's consequences are the main driver of norms. Taken together, these findings suggest that lying is not always perceived as blameworthy: some types of lies (such as those that benefit the receiver) are mostly viewed as acceptable and therefore rarely sanctioned. In this sense, the deontological view behind the concept of "pure" lying aversion applies only weakly to the punishment of false messages.⁴⁰

A potential explanation for why punishment for dishonesty is primarily driven by consequences is that legal sanctions are typically designed to deter behaviors that cause harm or generate gains at others' expense, such as negative externalities, rather than to punish unethical behavior *per se*. This logic is consistent with how fraud is defined in many legal systems. For example, under the UK legal definition of fraud, dishonesty alone is not sufficient: fraud also requires an intent to make a gain for oneself (or another), or to cause (or risk causing) a loss to another, in money or property terms.⁴¹ Hence, a beneficial lie would fall outside the core legal concept of fraud, as it involves deception without material loss to the receiver (or a third party).

Taken together with our evidence on reporting behavior, our results suggest that lying and truth-telling affect information transmission and its punishment in different ways. We show that lying aversion and truth-telling preferences play an asymmetric role in reporting behavior and punishment. This implies that, whenever possible, both preferences should be measured explicitly rather than inferring one from evidence on the other.

Our results also indicate that sanctions are more closely related to what people believe should be done (personal and injunctive judgments) than to what they expect others to do (descriptive norms). However, although the message's appropriateness accounts for an important part of the influence of dishonesty and truth-telling on punishment, its economic

⁴⁰"To be truthful (honest) in all declarations is therefore a sacred command of reason prescribing unconditionally, one not to be restricted by any conveniences." (Kant, 1797)

⁴¹See *Fraud Act 2006*, c. 35, ss. 2–5, available at <https://www.legislation.gov.uk/ukpga/2006/35>.

consequences retain an additional role. All in all, our results indicate that other-regarding preferences (especially concerns about inequality) are the main driver of punishment, with normative assessments also playing a substantial role.

The differences we observe in how dishonesty affects senders' reporting behavior and observers' punishment indicate that the determinants of sanctions can differ from those guiding people's behavior. This distinction can be particularly relevant from a policy perspective, especially when designing effective and sustainable behavioral interventions that rely on sanctions as norm enforcers. Tailoring sanctions to the consequences and social judgment of dishonest acts may help avoid backfiring effects, deterring spiteful communication without discouraging beneficial information transmission.

Although we document several forms of dishonest behavior, there remains scope to study punishment for other dishonest acts. Two relevant cases are harmful lies that benefit the sender at a small cost to the receiver, and altruistic lies that benefit the receiver at the sender's expense. A further step is to extend our information-transmission environment by allowing participants to remain silent, provide partial lies and truths, or communicate under imperfect information (Sánchez-Pagés and Vorsatz, 2007; Khalmetski, Rockenbach, and Werner, 2017), examining sanctions in these settings.

On the punishment side, an additional extension is to introduce uncertainty or probabilistic punishment, where a report is considered for sanction with an exogenous probability (possibly unknown). While there is evidence that the probability of punishment affects reporting behavior (Laske, Saccardo, and Gneezy, 2018), its effects on punishment remain largely unexplored. We view such an environment as a natural direction for future work due to its similarity to real examples of communication where sanctions play a crucial role, such as tax evasion.

A further natural extension is to introduce a compensation mechanism that complements sanctions. However, allowing for rewards may alter not

only reporting behavior, but also the perceived meaning of sanctioning dishonesty and truth-telling (see Chavez and Bicchieri (2013), Bicchieri and Maras (2022), and Civaï, Capraro, and Polonio (2025) for examples of how the introduction of rewards could alter punishment in other settings). Although it goes beyond the focus of this study, a framework where rewards and sanctions for communication coexist would be an interesting next step, complementing results in cooperative settings (Andreoni, Harbaugh, and Vesterlund, 2003).

Beyond the potential caveats related to these extensions, we acknowledge other limitations of our study. Despite the central role of economic consequences, we cannot disentangle how different other-regarding preferences shape punishment and communication. Besides, we cannot causally identify the role of self and social image concerns, which have been shown to influence dishonesty (Charness and Dufwenberg, 2006; Khalmetski and Sliwka, 2019; Bašić and Quercia, 2022). Extending our design to measuring these preferences more directly is an interesting avenue for future research.

Moreover, real-world communication environments involve repeated interactions, whereas our design focuses on one-shot interactions. Studying punishment in repeated information-transmission games would help assess how norm enforcement and the evolution of dishonest behavior shape sanctions over time. Finally, our online setting and predefined message space may affect communication and its sanctions, limiting external validity. Extensions of this study in laboratory or field settings, allowing for richer communication and interactions closer to everyday life, would be valuable.

Chapter 5

Conclusion

This dissertation advances the understanding of coordination, communication, and punishment as three distinct types of economic behavior. Individuals may coordinate on an action for a long time due to what happens in rare circumstances rather than in the most common scenario. A truthful message may be used strategically to mislead others, and senders may conceal their deceptive intentions when asked directly. Sanctions may appear to respond to dishonesty itself, while in fact being driven mainly by the consequences of the information transmitted. Using evolutionary game theory and experimental methods, this dissertation shows that identifying the mechanisms behind behavior is essential for understanding economic interactions.

Across the three essays, a common finding emerges: the most visible features of an interaction between individuals are not always reliable guides to the forces that shape behavior. Uncovering less visible mechanisms substantially changes both the interpretation of actions and the predictions derived from them, adding an essential component that goes beyond what decisions suggest at first glance.

The model of the first essay shows that the emergence and persistence of conventions under uncertainty cannot be understood by focusing only

on the most common or average environment. In coordination problems where payoffs evolve over time, the equilibria that emerge in the long run may depend on rare events and on how agents make mistakes. As a result, conventions may persist not because they are optimal in the environment that takes place most frequently, but because decisions made under very infrequent circumstances determine long-run stability.

Different underlying intentions may lead individuals to tell the truth or to lie when transmitting private information, so the second essay provides a new experimental method to identify deceptive behavior. Using actions in both a strategic and a non-strategic sender-receiver game together with elicited beliefs, this method distinguishes actual deceivers from pessimistic truth-tellers who did not aim to deceive despite expecting mistrust. Moreover, it captures senders who conceal their selfish intentions when asked about their expectations. These findings provide new evidence for the literature on dishonest behavior, showing that the share of individuals who aim to deceive may be higher than what previous approaches suggest.

A broader implication of this essay for the experimental literature is that relying only on questions about the sender's expected trust may be insufficient to identify senders' deceptive intentions correctly. Experimental economists should be cautious when inferring motives from stated beliefs about previous actions alone, especially when those answers may reveal sensitive information that participants prefer to hide.

These results motivate the last essay, which turns to one of the mechanisms that regulate communication: sanctions. The main finding is that punishment cannot be attributed solely to the veracity of the message, especially when the sender benefits economically from the transmitted information. The distributional implications of a message are crucial for understanding sanctions. Although truth-telling mitigates sanctions, individuals are punished when truth-telling has selfish consequences. By contrast, lying plays a greater role when consequential considerations are weaker, suggesting that its secondary effect is context-dependent. Together with the findings on pure and partial lying aversion obtained when

analyzing senders' behavior, this essay contributes to the literature on information transmission, showing that the two sides of communication may have different effects on punishment and reporting behavior.

Moreover, this essay shows that punishment reflects more closely what individuals believe ought to be done (both from a personal perspective and from what they believe society considers appropriate) than what they expect others to do. However, although personal and social norms capture a large share of the effect of message veracity, payoff consequences remain the main driver of sanctions. This indicates that distributional considerations operate as an additional force shaping punishment beyond normative judgments. These findings contribute not only to the literature on punishment and communication, but also to a broader line of research on the role of distributional considerations and social norms in economic behavior, by showing that both can jointly shape decisions through distinct and independent channels.

Despite its contributions to several fields, this dissertation also has certain limitations that point to directions for future research. The theoretical model developed in the first essay is based on a number of assumptions that limit its scope, such as focusing on myopic best-reply dynamics and restricting scenarios to two-action games. Allowing for alternative behavioral rules that coexist within the population, or extending the strategy set, would increase the generality of the results, although at a considerable cost in terms of tractability.

Furthermore, the identification method in the second essay applies only to settings where truth-telling and lying are the only two forms of communication. Extending the design to allow for a richer communication environment, in which senders may tell partial truths, partial lies, or remain silent, would be an interesting direction. Such an extension could help explain behaviors not captured in this dissertation and might also change the extent to which deceptive intentions can be identified. These limitations also apply to the experimental design of the third essay. For this essay, a further extension would be to introduce rewards as an alternative mechanism for encouraging truth-telling. A study in which

both rewards and sanctions coexist could help explain not only how individuals use both mechanisms, but also how they affect incentives for truthful communication.

Finally, a natural next step for this dissertation as a whole is to connect the insights across the three essays in a unified framework. One interesting avenue would be to study an information-transmission setting in which communication and trustworthiness coevolve in the presence of punishment. Such a framework would shed light on the emergence and evolution of truth-telling and dishonest behavior, and on how deceptive intentions and sanctioning behavior interact in the long run.

Although the three essays examine different forms of economic behavior, they share a common lesson: explaining human behavior often requires looking beneath the surface of actions. This dissertation shows that understanding the reasoning underlying decisions is essential for interpreting economic interactions. Uncovering these mechanisms remains a central challenge for economic research and a necessary step toward designing effective interventions that sustain coordination, promote honest communication, and support its effective and sustainable enforcement.

Appendix A

Appendix to Chapter 2

A.1 Proofs of statements in the main text

A.1.1 Proof of Proposition 1

Proof. We first show that convention A is an absorbing set. Consider two states, $\tilde{x} = (\tilde{\alpha}, \tilde{\mu}, \tilde{\sigma})$ and $x' = (\alpha', \mu', \sigma')$, $\tilde{x} \neq x'$, both belonging to convention A . Then, $\tilde{\alpha}_i = \alpha'_i = A \forall i \in N$. These two states communicate with each other since the probability of moving from $\tilde{x}$ to x' is at least as large as the probability that the Markov-switching process selects σ' , which is positive. The same argument applies in the opposite direction.

Now suppose that $\tilde{x}$ belongs to convention A , whereas x' does not. Then, $\alpha'_i = B$ for some agent $i \in N$. Since action A is a strict Nash equilibrium in every scenario of the game (considered in isolation), no agent receiving a revision opportunity at $\tilde{x}$ will deviate from A to B , since $BR_i(A) = A$ for all $i \in N$. Hence, x' cannot be reached from $\tilde{x}$ with positive probability. Therefore, convention A is an absorbing set. An analogous argument shows that convention B is also an absorbing set.

We next show that no other absorbing sets exist. Consider a state $\tilde{x}$ that does not belong to any of the aforementioned conventions, and suppose that at time t the process is in state $\tilde{x}$. Let $i \in N$ be an agent for whom $y \in \{A, B\}$ is a best response in the scenario played at $\tilde{x}$. Let Z denote the

set of agents other than i who choose an action different from y in state $\tilde{x}$.

With probability $\gamma(1 - \gamma)^{n-1}$, agent i is the only agent who receives a revision opportunity at time t , and chooses y by myopic best reply with positive probability (equal to 1 if y is the unique best response, and equal to $\frac{1}{2}$ otherwise). Note that agent i faces $|Z|$ opponents not choosing y . If $Z = \emptyset$, then a state in convention y has been reached.

If $Z \neq \emptyset$, then consider an agent $j \in Z$. Since M is ergodic, there exists an integer $m \geq 1$ such that, with positive probability, the scenario at period $t + m$ is again σ^t . In addition, with probability $(1 - \gamma)^{n(m-1)} > 0$, no agent receives a revision opportunity during periods $t+1, \dots, t+m-1$, so actions remain unchanged during the interim periods.

Then, with probability $\gamma(1 - \gamma)^{n-1}$, the only agent who receives a revision opportunity at time $t + m$ is agent j . Indeed, agent j faces a number of agents not choosing y equal to the cardinality of Z minus 1 since agent i chooses y and therefore does not belong to Z . Hence, with positive probability, at time $t + m$ the game is again in σ^t , and the only change in the action profile relative to time t is that agent i now chooses y , so agent j faces $|Z| - 1$ opponents not choosing y (one fewer than i did at time t). Hence, a fortiori, agent j chooses y under myopic best reply with positive probability, as done by agent i at time t .

As a result, the cardinality of Z gets reduced by one. If set Z becomes the empty set, a state in convention y has been reached. Otherwise, repeating the same argument, with positive probability, the set Z is cleared in a finite number of steps (at most the cardinality of Z , $|Z|$), and convention y is reached. Therefore, every state outside the two conventions can reach one of them with positive probability, so there are no other absorbing sets. $\square$

A.1.2 Proof of Proposition 2

Proof. Let's compute the radius and coradius of each convention in the whole game.

Denote as $\underline{R} = \min_k \{R_k(A)\}$. If $R(A) = \tilde{R} < \underline{R}$, then $\exists j \in \{1, \dots, \varsigma\} : R_j(A) = \tilde{R}$ (since all the agents play always in the same scenario), which contradicts the definition of minimum. Hence, $R(A) \geq \underline{R}$. Denote as $\underline{M} = \{G_k : R_k(A) = \underline{R} \text{ for } k = 1, \dots, \varsigma\}$. Then, if starting at convention A ,

$\underline{R}$ mistakes happen and the system switches to any scenario $G_k \in \underline{M}$, all the individuals will move to convention B (since the minimum number of mistakes to get out of A in this scenario has been reached), leaving convention A .

Therefore, we found paths where the minimum number of mistakes to get out of convention A is $\underline{R}$. Then, $R(A) = \underline{R}$. Analogously, $R(B) = \min\{R_1(B), \dots, R_\varsigma(B)\}$ where $R_k(B)$ is the radius of convention B in scenario G_k in isolation.

Denote as $\bar{R} = \min_k\{CR_1(A), \dots, CR_\varsigma(A)\}$ where $CR_k(A)$ is the coradius of convention A in scenario G_k in isolation. If $CR(A) = \hat{R} < \bar{R}$, then $\exists j \in \{1, \dots, \varsigma\} : CR_j(A) = \hat{R}$ (since all the agents play always in the same scenario), which contradicts the definition of minimum. Hence, $CR(A) \geq \bar{R}$. Denote as $\bar{M} = \{G_k : CR_k(A) = \bar{R} \text{ for } k = 1, \dots, \varsigma\}$. Then, starting at convention B , regardless of the scenario in which the game is placed, $\bar{R}$ mistakes may occur and the system may switch to any scenario $G_k \in \bar{M}$, in which all the individuals will move to convention A by best response, since the minimum number of mistakes to get into convention A in this scenario has been reached.

Therefore, we found paths where the minimum number of mistakes to get into convention A is $\bar{R}$ no matter the scenario in which the game starts (so we could start at the scenario where it is most difficult to switch from B to A). Then, $CR(A) = \bar{R}$. Using an analogous argument, $CR(B) = \min\{CR_1(B), \dots, CR_\varsigma(B)\}$ where $CR_k(B)$ is the coradius of convention B in scenario G_k in isolation.

□

A.1.3 Proof of Proposition 3

Proof. We just prove that $R(A) = \min_l\{n_l(B)\} \min_k\{r_k^B\}$ (analogous argument for $R(B)$).

First, let's prove that $R(A) \leq \min_l\{n_l(B)\} \min_k\{r_k^B\}$.

Let $\tilde{G}$ be one of the scenarios where the resistance to make mistakes by playing action B is minimum, i.e. $\tilde{G} \in \{G_q : r_q^B = \min_k\{r_k^B\}\}$. Let's also denote as $\tilde{G}$ one of the scenarios where the number of agents needed playing action B to switch the rest from playing A to B by best reply is minimum, i.e. $\tilde{G} \in \{G_m : n_m(B) = \min_l\{n_l(B)\}\}$.

We can construct a path of states starting at convention A and reaching convention B as follows. $n_{\bar{G}}(B)$ mistakes (from playing action A to B) are made being always placed in scenario $\bar{G}$ (so the resistance to mistakes is always $\min_k \{r_k^B\}$). Then, since M is ergodic, with positive probability the system changes to scenario $\bar{G}$ with no mistakes in the switching process. Hence, we are in a situation where $n_{\bar{G}}(B)$ agents are playing action B in scenario $\bar{G}$, so, by definition of $n_{\bar{G}}(B)$, the rest of agents switch to action B by myopic best reply, reaching then convention B .

Therefore, we constructed a path of states starting at convention A and ending in convention B , such that the sum of resistances of its states is $\min_l \{n_l(B)\} \min_k \{r_k^B\}$. Hence, by definition of radius, we have $R(A) \leq \min_l \{n_l(B)\} \min_k \{r_k^B\}$.

Now, let's prove that $R(A) \geq \min_l \{n_l(B)\} \min_k \{r_k^B\}$.

Suppose there exists a path of states from convention A to convention B , denoted as $\tilde{X}$, with a lower sum of resistances in this path, denoted as $\tilde{r} = \sum_{x \in \tilde{X}} r_x = \sum_{x \in \tilde{X}} r_x$ where $\tilde{X}$ is the set of states in $\tilde{X}$ with a positive resistance in the transition. Then,

$$\tilde{r} < \min_l \{n_l(B)\} \min_k \{r_k^B\} = \sum_{q=1}^{\min_l \{n_l(B)\}} \min_k \{r_k^B\}. \quad (\text{A.1})$$

By definition of minimum, since $\tilde{X}$ starts at convention A and ends at convention B ,

$$\tilde{r} = \sum_{x \in \tilde{X}} r_x \geq \sum_{x \in \tilde{X}} \min_k \{r_k^B\} = \sum_{s=1}^{|\tilde{X}|} \min_k \{r_k^B\} \quad (\text{A.2})$$

where $|\tilde{X}|$ is the cardinal of $\tilde{X}$, that is, the number of elements in $\tilde{X}$.

Combining equations (A.1) and (A.2), since $r_k^B > 0 \quad \forall k = 1, \dots, S$.

$$\sum_{x=1}^{|\tilde{X}|} \min_k \{r_k^B\} \leq \tilde{r} < \sum_{q=1}^{\min_l \{n_l(B)\}} \min_k \{r_k^B\},$$

This means that the number of mistakes (from playing action A to playing action B) made in the path $\tilde{X}$ to reach convention B is lower than $\min_l \{n_l(B)\}$.

Denote as $\check{G}$ the scenario in the path $\tilde{X}$ from which all remaining steps are done with zero resistance. Since $|\tilde{X}|$ mistakes from A to B were made in $\tilde{X}$, at this scenario convention B was reached with at least $|\tilde{X}|$ agents playing action B . Then, by definition of basin of attraction of convention B in the single scenario $\check{G}$, $n_{\check{G}}(B) \leq |\tilde{X}| < \min_l \{n_l(B)\}$, which contradicts the definition of minimum.

Therefore, $R(A) = \min_l \{n_l(B)\} \min_k \{r_k^B\}$.

□

A.1.4 Proof of Corollary 1

Proof. The proof is a particularization of the argument provided in Proposition 3. □

A.1.5 Proof of Proposition 4

Proof. We just prove the result for the radius of convention A , i.e., $R(A) = \min_k \{r(\pi_{G_k}^{AA})\} + \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\} \min_l \{n_l(B) - 1\}$. An analogous argument applies for the radius of convention B , $R(B)$.

First, let's prove that

$$R(A) \leq \min_k \{r(\pi_{G_k}^{AA})\} + \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\} \min_l \{n_l(B) - 1\}.$$

Denote as $\hat{G}$ one of the scenarios where the resistance to make mistakes if both agents play action A is minimum, i.e. $\hat{G} \in \{G_q : r(\pi_{G_q}^{AA}) = \min_k \{r(\pi_{G_k}^{AA})\}\}$ and $\tilde{G}$ one of the scenarios with the minimum resistance for an agent playing A no matter what the counterpart is playing, i.e. $\tilde{G} \in \{G_q : \min\{r(\pi_{G_q}^{AA}), r(\pi_{G_q}^{AB})\} = \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\}\}$. Denote as $\bar{G}$ one of the scenarios where the number of agents needed playing action B to switch the rest of players from playing A by best response to B is minimum, i.e. $\bar{G} \in \{G_m : n_m(B) = \min_l \{n_l(B)\}\}$.

We can construct a path of states starting at convention A such that the first mistake (from playing action A to B) is made in scenario $\tilde{G}$ (with resistance to mistakes $\min_k \{r(\pi_{G_k}^{AA})\}$), then $n_{\tilde{G}}(B) - 1$ mistakes are made being always placed in scenario $\tilde{G}$, (so the resistance to these mistakes is always $\min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\}$), and then the system changes to scenario $\bar{G}$ with no mistakes in the switching process (which happens with positive probability since M is ergodic). Hence, we are in a situation where $n_{\bar{G}}(B)$ agents are playing action B in scenario $\bar{G}$ (one of them changed at scenario $\tilde{G}$ and the rest at $\bar{G}$). Then, by definition of $n_{\bar{G}}(B)$, the rest of agents switch to action B by best reply, reaching convention B .

Therefore, we constructed a path of states from convention A to convention B such that the sum of resistance of its states is $\min_k \{r(\pi_{G_k}^{AA})\} + \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\} \min_l \{n_l(B) - 1\}$. So, by definition of radius, $R(A) \leq \min_k \{r(\pi_{G_k}^{AA})\} + \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\} \min_l \{n_l(B) - 1\}$.

Now, let's prove that

$$R(A) \geq \min_k \{r(\pi_{G_k}^{AA})\} + \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\} \min_l \{n_l(B) - 1\}.$$

Denote as $\tilde{X}$ the path of states from convention A to convention B with the lowest sum of resistances, denoted as $\tilde{r} = \sum_{x \in \tilde{X}} r_x = \sum_{x \in \bar{X}} r_x$ where $\bar{X}$ is the set of states in $\tilde{X}$ with a positive resistance in the transition.

Suppose that

$$\begin{aligned} \tilde{r} &< \min_k \{r(\pi_{G_k}^{AA})\} + \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\} \min_l \{n_l(B) - 1\} \\ &= \min_k \{r(\pi_{G_k}^{AA})\} + \sum_{q=1}^{\min_l \{n_l(B)-1\}} \min_k \{\min\{r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB})\}\}. \end{aligned} \quad (\text{A.3})$$

Denote the initial state in $\tilde{X}$ as $\tilde{x}^0 = (\alpha^0, \mu^0, G^0)$. Since $\tilde{X}$ starts at convention A , $\alpha_i^0 = A \forall i \in N$. Hence, the resistance of the first mistake is $r_{x_0} = r(i, x^0) = r(\pi_{G^0}^{AA}) \forall i \in N$.

Then, by definition of minimum

$$r_{x_0} = r(\pi_{G^0}^{AA}) \geq \min_k \{r(\pi_{G_k}^{AA})\}. \quad (\text{A.4})$$

Moreover, since $\tilde{X}$ is a path with minimum resistance that goes from

convention A to convention B , agents will only make mistakes by playing action B instead of A , but not in the other way around. Indeed, a mistake from B to A has positive resistance and decreases the number of agents currently playing B , so it cannot help the process move toward convention B . Hence any minimum-resistance path containing such a mistake can be replaced by another path from convention A to convention B with weakly lower total resistance in which that mistake is omitted. Hence, the only transitions in path $\tilde{X}$ with positive resistance are those where agents who made the mistake switch from A to B , either against an agent who plays action A or against one playing action B , depending on what payoff is minimum. Then,

$$\begin{aligned} \sum_{\substack{x \in \tilde{X} \\ x \neq x_0}} r_x &\geq \sum_{\substack{x \in \tilde{X} \\ x \neq x_0}} \min_k \{ \min \{ r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB}) \} \} \\ &= \sum_{s=1}^{|\tilde{X}|-1} \min_k \{ \min \{ r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB}) \} \} \end{aligned} \quad (\text{A.5})$$

Then, joining the results in equations (A.4) and (A.5)

$$\begin{aligned} \tilde{r} &= \sum_{x \in \tilde{X}} r_x = r_{x_0} + \sum_{\substack{x \in \tilde{X} \\ x \neq x_0}} r_x \\ &\geq \min_k \{ r(\pi_{G_k}^{AA}) \} + \sum_{s=1}^{|\tilde{X}|-1} \min_k \{ \min \{ r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB}) \} \} \end{aligned} \quad (\text{A.6})$$

Hence, by combining equations (A.3) and (A.6), we obtain $|\tilde{X}| - 1 < \min_l \{ n_l(B) - 1 \}$, which implies that $|\tilde{X}| < \min_l \{ n_l(B) \}$. This means that the number of mistakes (from playing action A to playing action B) made in the path $\tilde{X}$ to reach convention B is lower than $\min_l \{ n_l(B) \}$.

Denote as $\check{G}$ the scenario in the path $\tilde{X}$ from which all remaining steps are done with zero resistance. Since only $|\tilde{X}|$ mistakes from A to B where made in $\tilde{X}$, at this scenario convention B was reached with only $|\tilde{X}|$ agents playing action B . Then, by definition of basin of attraction of convention B in the single scenario $\check{G}$, $n_{\check{G}}(B) \leq |\tilde{X}| < \min_l \{ n_l(B) \}$, which contradicts the definition of minimum.

Therefore,

$$R(A) = \min_k \{ r(\pi_{G_k}^{AA}) \} + \min_k \{ \min \{ r(\pi_{G_k}^{AA}), r(\pi_{G_k}^{AB}) \} \} \min_l \{ n_l(B) - 1 \}.$$

□

A.1.6 Proof of Proposition 5

Proof. We just prove that

$$R(A) = \sum_{j=0}^{\min_l \{n_l(B)-1\}} \min_k \left\{ r_k^B \left[\frac{n-1-j}{n-1} (\pi_{G_k}^{AA} - \pi_{G_k}^{BA}) + \frac{j}{n-1} (\pi_{G_k}^{AB} - \pi_{G_k}^{BB}) \right] \right\}.$$

First, let's prove that

$$R(A) \leq \sum_{j=0}^{\min_l \{n_l(B)-1\}} \min_k \left\{ r_k^B \left[\frac{n-1-j}{n-1} (\pi_{G_k}^{AA} - \pi_{G_k}^{BA}) + \frac{j}{n-1} (\pi_{G_k}^{AB} - \pi_{G_k}^{BB}) \right] \right\}.$$

Denote as $\Pi_k^D(j)$ the expected payoff in scenario G_k of playing action D when j agents were playing action B and the remain $n-1-j$ agents action A , then

$$\Pi_k^A(j) = \frac{n-1-j}{n-1} \pi_{G_k}^{AA} + \frac{j}{n-1} \pi_{G_k}^{AB}$$

$$\Pi_k^B(j) = \frac{n-1-j}{n-1} \pi_{G_k}^{BA} + \frac{j}{n-1} \pi_{G_k}^{BB}$$

where $\pi_{G_k}^{\alpha_i \alpha_j}$ is agent i 's payoff of playing action α_i against an agent who plays action α_j in scenario G_k .

Denote as $\tilde{G}_j$ one of the scenarios where the difference of expected payoff by playing action A instead of action B when j agents play action B is minimum, i.e., $\tilde{G}_j \in \{G_q : r_q^B(i, x) = \min_k \{r_k^B(\Pi_k^A(j) - \Pi_k^B(j))\}$, where j is the number of agents who play B in x .

Denote as $\bar{G}$ one of the scenarios with the minimum basin of attraction of convention B , that is, where the number of agents needed playing action B to switch the rest from playing action A to B is minimum, i.e. $\bar{G} \in \{G_m : n_m(B) = \min_l \{n_l(B)\}\}$.

We can construct a path of states starting at convention A and scenario $\tilde{G}_0$ such that when only j agents play action B , a mistake is made at scenario $\tilde{G}_j$, and right after that mistake the system changes to scenario $\tilde{G}_{j+1}$. This process will be done until $n_{\bar{G}}(B)$ agents have made a mistake (from choosing A to B). Then, with positive probability, the system

changes to scenario $\bar{G}$ with no mistakes in the switching process, reaching a state where $n_{\bar{G}}(B)$ agents play action B in scenario $\bar{G}$ (those who made a mistake in the process). Then, by definition of $n_{\bar{G}}(B)$, the rest of agents switch to action B (since it is the best response in this scenario), reaching then convention B .

In the states of the path before the last one, where less than $n_{\bar{G}}(B)$ mistakes have been made, less than $n_{\bar{G}}(B)$ agents are playing action B , so it is not possible that the agents who play A prefer to switch to B since $n_{\bar{G}}(B) = \min_l \{n_l(B)\}$. Then, A is an expected best response, which implies that, for all $k = 1, \dots, S$,

$$\frac{n-1-j}{n-1}(\pi_{G_k}^{AA} - \pi_{G_k}^{BA}) + \frac{j}{n-1}(\pi_{G_k}^{AB} - \pi_{G_k}^{BB}) > 0, \quad \forall j = 0, \dots, n_{\bar{G}}(B) - 1.$$

Therefore, we constructed a path of states from convention A to convention B such that the sum its resistances, which we denote as $\bar{r}$, is

$$\begin{aligned} \bar{r} &= \sum_{j=0}^{\min_l \{n_l(B)-1\}} \min_k \left\{ r_k^B \left[\Pi_k^A(j) - \Pi_k^B(j) \right] \right\} \\ &= \sum_{j=0}^{\min_l \{n_l(B)-1\}} \min_k \left\{ r_k^B \left[\frac{n-1-j}{n-1}(\pi_{G_k}^{AA} - \pi_{G_k}^{BA}) + \frac{j}{n-1}(\pi_{G_k}^{AB} - \pi_{G_k}^{BB}) \right] \right\}. \end{aligned}$$

Hence, by definition of radius,

$$R(A) \leq \sum_{j=0}^{\min_l \{n_l(B)-1\}} \min_k \left\{ r_k^B \left[\frac{n-1-j}{n-1}(\pi_{G_k}^{AA} - \pi_{G_k}^{BA}) + \frac{j}{n-1}(\pi_{G_k}^{AB} - \pi_{G_k}^{BB}) \right] \right\}.$$

Now, let's prove that

$$R(A) \geq \sum_{j=0}^{\min_l \{n_l(B)-1\}} \min_k \left\{ r_k^B \left[\frac{n-1-j}{n-1}(\pi_{G_k}^{AA} - \pi_{G_k}^{BA}) + \frac{j}{n-1}(\pi_{G_k}^{AB} - \pi_{G_k}^{BB}) \right] \right\}.$$

By definition, $n_k(B)$ is the minimum number of agents playing B in scenario G_k such that B becomes a best reply for the remaining agents at scenario G_k . Then, it is not possible that at scenario G_k , B becomes a best response with less than $n_k(B)$ agents playing B . Hence, the expected payoff of playing A is higher than the one of playing B at scenario G_k ,

that is, for every scenario G_k ,

$$\Pi_k^A(j) - \Pi_k^B(j) > 0, \forall j = 0, \dots, n_k(B) - 1.$$

If $\tilde{G} \in \{G_m : n_m(B) = \min_l \{n_l(B)\}\}$, then by definition of minimum, $n_{\tilde{G}}(B) \leq n_k(B) \forall k = 1, \dots, S$. Hence, for every scenario G_k , $\Pi_k^A(j) - \Pi_k^B(j) > 0 \forall j = 0, \dots, n_{\tilde{G}}(B) - 1$. So, in all the scenarios, with a lower number of agents playing action B than $n_{\tilde{G}}(B)$, there is no gain by deviating to play action B .

If there is a path with minimum sum of resistances such that the number of steps with a positive resistance is higher than $n_{\tilde{G}}(B)$, then, we could create a path such that it is similar until you reach $n_{\tilde{G}}(B)$ transitions with positive resistance, and then the system switches to $\tilde{G}$ with positive probability and no mistakes in the switching process where, by definition of basin of attraction, convention B is reached, having this new path lower sum of resistances, which contradicts the definition of minimum.

Denote as $\tilde{X}$ the path with minimum resistance that connects conventions A and B . Hence, $\nu \leq n_{\tilde{G}}(B)$ where ν is the number of steps with a positive resistance in $\tilde{X}$. We now prove that $\nu \geq n_{\tilde{G}}(B)$.

Since for every scenario G_k , $\Pi_k^A(j) - \Pi_k^B(j) > 0 \forall j = 0, \dots, n_{\tilde{G}}(B) - 1$, action B is never a best response for any player when less than $n_{\tilde{G}}(B)$ agents play action B . Then, the increase in the number of agents who play B in the transitions with less than $n_{\tilde{G}}(B)$ steps when starting from convention A is only driven by mistakes.

Hence, if $\nu < n_{\tilde{G}}(B)$, in the last scenario of $\tilde{X}$, denote it as $\tilde{G}$, we have that $n - \nu$ agents were playing action A and then convention B was reached with only ν agents playing B . Then $n_{\tilde{G}}(B) = \nu < n_{\tilde{G}}(B) = \min_l \{n_l(B)\}$, which contradicts the definition of minimum.

Therefore, the path with minimum resistance that connects conventions A and B must have $n_{\tilde{G}}(B) = \min_l \{n_l(B)\}$ transitions with positive resistance, which is equal to the number of mistakes made along the path.

We can express the resistance of the path $\tilde{X}$, denoted as $\tilde{r}$, as

$$\tilde{r} = \sum_{x \in \tilde{X}} r_x^B = \sum_{x \in \tilde{X}} r_x^A$$

where $\bar{X}$ is the set of states in $\tilde{X}$ where the resistance to mistakes in the transition is positive, so $|\bar{X}| = n_{\bar{G}}(B) = \min_l \{n_l(B)\}$.

Moreover, all transitions but the last one in $\tilde{X}$ must be either between states where the scenario is switched while preserving all the players the same action (transitions with zero resistance), or where one agent makes a mistake remaining everybody else and the scenario played equal (transitions with positive resistance).

Then, the initial and terminal states of each transition in $\bar{X}$ have the same scenario and all contain a different number of agents playing action B (always smaller than $\min_l \{n_l(B)\}$).

Therefore, if

$$\tilde{r} < \sum_{j=0}^{\min_l \{n_l(B)-1\}} \min_k \{r_k^B [\Pi_k^A(j) - \Pi_k^B(j)]\},$$

then, there exists $s' = (\alpha_{s'}, \mu_{s'}, G_{s'}) \in \bar{X}$ such that

$$r_{s'}^B [\Pi_{s'}^A(j') - \Pi_{s'}^B(j')] < \min_k \{r_k^B [\Pi_k^A(j') - \Pi_k^B(j')]\},$$

where $r_{s'}$ is the resistance function at scenario s' and j' is the number of players in $\alpha_{s'}$ who play action B . This contradicts the definition of minimum.

Therefore,

$$R(A) = \sum_{j=0}^{\min_l \{n_l(B)-1\}} \min_k \left\{ r_k^B \left[\frac{n-1-j}{n-1} (\pi_{G_k}^{AA} - \pi_{G_k}^{BA}) + \frac{j}{n-1} (\pi_{G_k}^{AB} - \pi_{G_k}^{BB}) \right] \right\}.$$

□

A.1.7 Proof of Proposition 6

Proof. The proof consists of a mere observation. Any path between absorbing sets involves transitions that entail changes in actions and, possibly, changes in scenarios. It is enough to note that, if T' and T'' are ergodic, then any transition between scenarios occurs in a finite number of steps

with a positive probability that remains bounded away from zero as the noise becomes vanishingly small, and hence it adds no cost to the overall resistance of the path between the associated absorbing sets. $\square$

A.2 Examples

A.2.1 A paradigmatic example

This example illustrates the crucial role of the error model in determining the pivotal scenario for computing radii and coradii in this setting.

Consider a game G consisting of the scenarios $G_1, \dots, G_6$, described below:

	A	B
A	10,10	0,0
B	0,0	1,1

Scenario G_1

	A	B
A	2,2	-18,0
B	0,-18	3,3

Scenario G_2

	A	B
A	0.5,0.5	0,0.1
B	0.1,0	0.5,0.5

Scenario G_3

	A	B
A	1,1	2,-2
B	-2,2	20,20

Scenario G_4

	A	B
A	1,1	-50,-30
B	-30,-50	2,2

Scenario G_5

	A	B
A	1,1	-22,-27
B	-27,-22	7,7

Scenario G_6

where the associated Markov process M that rules the transitions among $G_1, \dots, G_6$ has the following transition matrix:

$$T = \begin{bmatrix} 0.99 & 0.002 & 0.002 & 0.002 & 0.002 & 0.002 \\ 0.99 & 0.01 & 0 & 0 & 0 & 0 \\ 0.99 & 0 & 0.01 & 0 & 0 & 0 \\ 0.99 & 0 & 0 & 0.01 & 0 & 0 \\ 0.99 & 0 & 0 & 0 & 0.01 & 0 \\ 0.99 & 0 & 0 & 0 & 0 & 0.01 \end{bmatrix}, \quad (\text{A.7})$$

so M is ergodic but induces the system to stay approximately 99% of the time in G_1 .

Assume a population large enough, for instance, of size $n = 1000$. At first glance, one may be tempted to think that convention A is the one where the system will spend most time in the long run since G_1 is by far the most likely scenario and $f_{G_1}^A = 1/11 < 10/11 = f_{G_1}^B$, i.e., in G_1 the basin of attraction of A is ten times the basin of attraction of B .

However, due to the presence of the other five scenarios, convention B turns out to be always the only stochastically stable convention. Besides, it turns out that the scenario which is essential to make convention B stochastically stable depends crucially on the error model considered.

In the case of uniform mistakes, since the likelihood to make a mistake is independent of the current state of the system, to compute the radius of each convention is enough to take the complement to one of the basins of attraction of the other convention. See Table 12 for the relevant calculations.

Scenario	For R(A)	For R(B)
G_1	$f_{G_1}^B = 10/11 \simeq 0.9091$	$f_{G_1}^A = 1/11 \simeq 0.0909$
G_2	$f_{G_2}^B = 2/23 \simeq 0.0870$	$f_{G_2}^A = 21/23 \simeq 0.9130$
G_3	$f_{G_3}^B = 4/9 \simeq 0.4444$	$f_{G_3}^A = 5/9 \simeq 0.5556$
G_4	$f_{G_4}^B = 1/7 \simeq 0.1429$	$f_{G_4}^A = 6/7 \simeq 0.8571$
G_5	$f_{G_5}^B = 31/83 \simeq 0.3735$	$f_{G_5}^A = 52/83 \simeq 0.6265$
G_6	$f_{G_6}^B = 28/57 \simeq 0.4912$	$f_{G_6}^A = 29/57 \simeq 0.5088$

Table 12: Scenario characteristics that are relevant under uniform mistakes: $f_{G_k}^Z$ is the minimum fraction of agents needed to play action $Z \in \{A, B\}$ to make Z a best response in scenario G_k

Applying Proposition 2, we immediately get that $R(B) = \lceil \frac{1}{11}n \rceil > \lceil \frac{2}{23}n \rceil = R(A)$, implying that the only stochastically stable convention is convention B . This follows from the fact that $\min_k \{f_{G_k}^B\} = f_{G_2}^B = \frac{2}{23} < \frac{1}{11} = f_{G_1}^A = \min_k \{f_{G_k}^A\}$ which, given $n = 1000$, implies that

$\min_k \{n_k(B)\} < \min_k \{n_k(A)\}$, where $n_k(Z) = \lceil f_{G_k}^Z n \rceil$ is the minimum number of agents needed to play action $Z \in \{A, B\}$ in order to make Z a best response in scenario G_k .

For any scenario G_s , let $G^{(-s)}$ denote the reduced game obtained from G by deleting scenario G_s , and let $T^{(-s)}$ denote the corresponding transition matrix, obtained by deleting the s^{th} row and column of T and adjusting the transition probability of G_1 so that rows sum to one. Interestingly, if we consider the reduced game $G^{(-2)}$, obtained from G by deleting scenario G_2 , and let $T^{(-2)}$ be the associated transition matrix, obtained by deleting the second row and column of T and setting $T_{G_1, G_1}^{(-2)} = 0.992$, then applying again Proposition 2, we get that $R(A) = \lceil \frac{1}{7}n \rceil > \lceil \frac{1}{11}n \rceil = R(B)$ since $\min_{k \neq 2} \{f_{G_k}^A\} = f_{G_1}^A = \frac{1}{11} < \frac{1}{7} = f_{G_4}^B = \min_{k \neq 2} \{f_{G_k}^B\}$ and $n = 1000$. Hence, convention A turns out to be the unique stochastically stable convention in $G^{(-2)}$.

In the case of scenario/action-dependent mistakes, the likelihood to make a mistake depends possibly on the current scenario and the action played. Many different specifications of this error model are possible. To fix ideas, consider a specification such that $r_{G_6}^Z < r_{G_1}^Z < r_{G_k}^Z$, $Z \in \{A, B\}$, $k \in \{2, \dots, 5\}$, and, at the same time, $r_{G_6}^A = 5 > 1 = r_{G_6}^B$, $r_{G_1}^B = 20 > 10 = r_{G_1}^A$. Then, applying Proposition 3, we immediately get that $R(B) = 5 \lceil \frac{1}{11}n \rceil > \lceil \frac{2}{23}n \rceil = R(A)$ and, therefore, that B is the unique stochastically stable convention. This follows from the fact that the resistance to mistakes is least in scenario G_6 while playing A together with the fact that, for $n = 1000$, $\min_k \{n_k(B)\} = \min_k \lceil f_{G_k}^B n \rceil = \lceil \frac{2}{23}n \rceil < \lceil \frac{1}{11}n \rceil = \min_k \lceil f_{G_k}^A n \rceil = \min_k \{n_{G_k}(A)\}$ (see Table 12 for the values of $f_{G_k}^Z$, $Z \in \{A, B\}$).

Interestingly, if we consider the reduced game $G^{(-6)}$, obtained from G by deleting scenario G_6 , and let $T^{(-6)}$ be the associated transition matrix, obtained by deleting the sixth row and column of T and setting $T_{G_1, G_1}^{(-6)} = 0.992$, then applying again Proposition 3, we obtain that $R(A) = 20 \lceil \frac{2}{23}n \rceil > 10 \lceil \frac{1}{11}n \rceil = R(B)$ and, hence, that convention A is the unique stochastically stable convention in $G^{(-6)}$. This happens because, in the absence of G_6 , the resistance to mistakes is least in G_1 while playing B and this more than compensates the fact that, for $n = 1000$, $\min_{k \neq 6} \{n_k(B)\} =$

$$\min_{k \neq 6} \lceil f_{G_k}^B n \rceil = \lceil \frac{2}{23} n \rceil < \lceil \frac{1}{11} n \rceil = \min_{k \neq 6} \lceil f_{G_k}^A n \rceil = \min_{k \neq 6} \{n_k(A)\}.$$

An interesting specification of scenario/action-dependent mistakes is the case of intentional mistakes, where agents only make mistakes only when they can lead to a preferred equilibrium in the current scenario, and are more likely to do so if the payoff gain of switching to the other equilibrium is greater. Table 13 reports the payoff differences between conventions A and B for each scenario.

Scenario	For R(A)	For R(B)
G_1	$\pi_{G_1}^{BB} - \pi_{G_1}^{AA} = -9$	$\pi_{G_1}^{AA} - \pi_{G_1}^{BB} = 9$
G_2	$\pi_{G_2}^{BB} - \pi_{G_2}^{AA} = 1$	$\pi_{G_2}^{AA} - \pi_{G_2}^{BB} = -1$
G_3	$\pi_{G_3}^{BB} - \pi_{G_3}^{AA} = 0$	$\pi_{G_3}^{AA} - \pi_{G_3}^{BB} = 0$
G_4	$\pi_{G_4}^{BB} - \pi_{G_4}^{AA} = 19$	$\pi_{G_4}^{AA} - \pi_{G_4}^{BB} = -19$
G_5	$\pi_{G_5}^{BB} - \pi_{G_5}^{AA} = 1$	$\pi_{G_5}^{AA} - \pi_{G_5}^{BB} = -1$
G_6	$\pi_{G_6}^{BB} - \pi_{G_6}^{AA} = 6$	$\pi_{G_6}^{AA} - \pi_{G_6}^{BB} = -6$

Table 13: Payoff differences between convention A and convention B in each scenario that are relevant under intentional mistakes.

For simplicity, we consider the case in which the resistance to mistakes is equal to the reciprocal of the payoff difference between conventions when the difference is positive, and equal to 0 otherwise. Applying Corollary 1, we get that $R(B) = \frac{1}{9} \lceil \frac{1}{11} n \rceil > \frac{1}{19} \lceil \frac{2}{23} n \rceil = R(A)$ and, hence, that convention B is the unique stochastically stable convention. This follows from the fact that $\max_k \{\pi_{G_k}^{BB} - \pi_{G_k}^{AA}\} = \pi_{G_4}^{BB} - \pi_{G_4}^{AA} = 19 > 9 = \pi_{G_1}^{AA} - \pi_{G_1}^{BB} = \max_k \{\pi_{G_k}^{AA} - \pi_{G_k}^{BB}\}$, which implies that $\min_k \{r_k(\pi^{BB} - \pi^{AA})\} < \min_k \{r_k(\pi^{AA} - \pi^{BB})\}$, and also considering that, for $n = 1000$, $\min_k \{n_k(B)\} = \lceil \frac{2}{23} n \rceil < \lceil \frac{1}{11} n \rceil = \min_k \{n_k(A)\}$.

However, also in this case the result changes drastically by eliminating one scenario. Consider the reduced game $G^{(-4)}$, obtained from G by deleting scenario G_4 , and let $T^{(-4)}$ be the associated transition matrix, obtained by deleting the fourth row and column of T and setting $T_{G_1, G_1}^{(-4)} = 0.992$. By applying again Corollary 1, we see that $R(A) = \frac{1}{6} \lceil \frac{2}{23} n \rceil > \frac{1}{9} \lceil \frac{1}{11} n \rceil = R(B)$ and, hence, that convention A is the unique

stochastically stable convention in $G^{(-4)}$. This follows from the fact that, in this reduced game, $\max_{k \neq 4} \{\pi_{G_k}^{BB} - \pi_{G_k}^{AA}\} = \pi_{G_6}^{BB} - \pi_{G_6}^{AA} = 6 < 9 = \pi_{G_1}^{AA} - \pi_{G_1}^{BB} = \max_{k \neq 4} \{\pi_{G_k}^{AA} - \pi_{G_k}^{BB}\}$, which more than compensates the fact that, for $n = 1000$, $\min_{k \neq 4} \{n_k(B)\} = \min_{k \neq 4} \lceil f_{G_k}^B n \rceil = \lceil \frac{2}{23} n \rceil < \lceil \frac{1}{11} n \rceil = \min_{k \neq 4} \lceil f_{G_k}^A n \rceil = \min_{k \neq 4} \{n_k(A)\}$.

In the case of condition-dependent mistakes, what determines the radius of a convention is, besides its basin of attraction, the minimum of the possible payoffs an agent can earn by playing the correspondent action (provided that the population is large enough so that the first mistake, which is necessarily made after having earned the coordination payoff, is negligible). Table 14 reports the relevant payoffs for this example.

	For R(A)	For R(B)
G_1	$\pi_{G_1}^{AA} = 10 \mid \min\{\pi_{G_1}^{AA}, \pi_{G_1}^{AB}\} = 0$	$\pi_{G_1}^{BB} = 1 \mid \min\{\pi_{G_1}^{BA}, \pi_{G_1}^{BB}\} = 0$
G_2	$\pi_{G_2}^{AA} = 2 \mid \min\{\pi_{G_2}^{AA}, \pi_{G_2}^{AB}\} = -18$	$\pi_{G_2}^{BB} = 3 \mid \min\{\pi_{G_2}^{BA}, \pi_{G_2}^{BB}\} = 0$
G_3	$\pi_{G_3}^{AA} = 0.5 \mid \min\{\pi_{G_3}^{AA}, \pi_{G_3}^{AB}\} = 0$	$\pi_{G_3}^{BB} = 0.5 \mid \min\{\pi_{G_3}^{BA}, \pi_{G_3}^{BB}\} = 0.1$
G_4	$\pi_{G_4}^{AA} = 1 \mid \min\{\pi_{G_4}^{AA}, \pi_{G_4}^{AB}\} = 1$	$\pi_{G_4}^{BB} = 20 \mid \min\{\pi_{G_4}^{BA}, \pi_{G_4}^{BB}\} = -2$
G_5	$\pi_{G_5}^{AA} = 1 \mid \min\{\pi_{G_5}^{AA}, \pi_{G_5}^{AB}\} = -50$	$\pi_{G_5}^{BB} = 2 \mid \min\{\pi_{G_5}^{BA}, \pi_{G_5}^{BB}\} = -30$
G_6	$\pi_{G_6}^{AA} = 1 \mid \min\{\pi_{G_6}^{AA}, \pi_{G_6}^{AB}\} = -22$	$\pi_{G_6}^{BB} = 7 \mid \min\{\pi_{G_6}^{BA}, \pi_{G_6}^{BB}\} = -27$

Table 14: Payoffs earned in each scenario which are relevant for the case of condition-dependent mistakes.

Here we consider the case where resistance to mistakes is given by the exponential of the payoff earned. Applying Proposition 4, we obtain that $R(B) = e^{0.5} + \frac{1}{e^{30}} (\lceil \frac{1}{11} n \rceil - 1) > e^{0.5} + \frac{1}{e^{50}} (\lceil \frac{2}{23} n \rceil - 1) = R(A)$ and, hence, convention B is the only one stochastically stable under condition-dependent mistakes.

This follows intuitively from three considerations. First, $\min_k \{\pi_{G_k}^{BB}\} = 0.5 = \min_k \{\pi_{G_k}^{AA}\}$, so the minimum resistance associated with the first mistake is the same under both conventions. Second, $\min_k \{\min\{\pi_{G_k}^{BB}, \pi_{G_k}^{BA}\}\} = \pi_{G_5}^{BA} = -30 > -50 = \pi_{G_5}^{AB} = \min_k \{\min\{\pi_{G_k}^{AA}, \pi_{G_k}^{AB}\}\}$, so, for subsequent mistakes, the minimum resistance is associated with switching from A to

B. Third, for $n = 1000$, $\min_k \{n_k(B)\} = \min_k \lceil f_{G_k}^B n \rceil = \lceil \frac{2}{23} n \rceil < \lceil \frac{1}{11} n \rceil = \min_k \lceil f_{G_k}^A n \rceil = \min_k \{n_{G_k}(A)\}$, which implies that convention *B* has the larger basin of attraction.

However, also in this case the result changes drastically by eliminating one scenario. Consider the reduced game $G^{(-5)}$, obtained from G by deleting scenario G_5 , and let $T^{(-5)}$ be the associated transition matrix, obtained by deleting the fifth row and column of T and setting $T_{G_1, G_1}^{(-5)} = 0.992$.

By applying again Proposition 4, we obtain that

$$R(A) = e^{0.5} + \frac{1}{e^{22}} \left(\left\lceil \frac{2}{23} n \right\rceil - 1 \right) > e^{0.5} + \frac{1}{e^{27}} \left(\left\lceil \frac{1}{11} n \right\rceil - 1 \right) = R(B).$$

Therefore, convention *A* is the unique stochastically stable convention in $G^{(-5)}$. This follows from $\min_{k \neq 5} \{\pi_{G_k}^{BB}\} = 0.5 = \min_{k \neq 5} \{\pi_{G_k}^{AA}\}$, as seen for the case above, and $\min_{k \neq 5} \{\min\{\pi_{G_k}^{BB}, \pi_{G_k}^{BA}\}\} = \pi_{G_6}^{BA} = -27 < -22 = \pi_{G_6}^{AB} = \min_{k \neq 5} \{\min\{\pi_{G_k}^{AA}, \pi_{G_k}^{AB}\}\}$, which compensates the fact that, for $n = 1000$, $\min_k \{n_k(B)\} = \min_{k \neq 5} \lceil f_{G_k}^B n \rceil = \lceil \frac{2}{23} n \rceil < \lceil \frac{1}{11} n \rceil = \min_{k \neq 5} \lceil f_{G_k}^A n \rceil = \min_{k \neq 5} \{n_k(A)\}$.

In the case of cost-dependent mistakes, what determines the radius of a convention is, besides its basin of attraction, the expected loss associated with the action played by mistake. Table 15 reports the expected payoff losses for each scenario as functions of the fraction of the population playing action *A* or *B*.

In order to apply Proposition 5, it is useful to identify preliminarily what are the relevant scenarios for the calculation of the radii. In this example, it turns out that 2 different scenarios are important for the calculation of each radius. Let $\phi(A)$ and $\phi(B)$ be the fraction of agents playing *A* and *B* respectively. For $\phi_B \in [0, \frac{16}{221}]$, the lowest expected payoff loss associated with making a mistake that leads to play *B* is attained in G_3 . For $\phi_B \in [\frac{16}{221}, f_{G_2}^B)$, that is, $\phi_B \in [\frac{16}{221}, \frac{2}{23})$, the lowest expected payoff loss associated with making a mistake that leads to play *B* is instead attained in G_2 . Moreover, for $\phi_A \in [0, \frac{5}{101}]$ the lowest expected

Scenario	Payoff loss function	
	For R(A)	For R(B)
G_1	$10 - 11 \phi_B$	$1 - 11 \phi_A$
G_2	$2 - 23 \phi_B$	$21 - 23 \phi_A$
G_3	$0.4 - 0.9 \phi_B$	$0.5 - 0.9 \phi_A$
G_4	$3 - 21 \phi_B$	$18 - 21 \phi_A$
G_5	$31 - 83 \phi_B$	$52 - 83 \phi_A$
G_6	$28 - 57 \phi_B$	$29 - 57 \phi_A$

Table 15: Relevant payoff losses at each scenario under cost-dependent mistakes. ϕ_A (ϕ_B) is the fraction of agents who play action A (B) in the whole population.

payoff loss associated with making a mistake that leads to play A is attained in G_3 , while for $\phi_A \in [\frac{5}{101}, f_{G_1}^A)$, that is, $\phi_A \in [\frac{5}{101}, \frac{1}{11})$, the expected payoff loss associated with making a mistake that leads to play A is attained in G_1 .

Note that we can apply Proposition 5 without the necessity to consider the cases where $\phi_B \geq f_{G_2}^B$ for the calculation of $R(A)$, and $\phi_A \geq f_{G_1}^A$ for the calculation of $R(B)$, since at those threshold values no more mistakes are necessary to trigger a transition from one convention to the other.

Lastly, for simplicity we assume that the resistance to make a mistake is given by the exponential of the absolute value of the expected payoff loss. Hence, applying Proposition 5, we obtain that:

$$\begin{aligned}
 R(B) &= \sum_{j=0}^{\lceil \frac{5}{101} n \rceil - 1} e^{0.5 - 0.9 \frac{j}{n-1}} + \sum_{j=\lceil \frac{5}{101} n \rceil}^{\lceil \frac{1}{11} n \rceil - 1} e^{1 - 11 \frac{j}{n-1}} \\
 &> \sum_{j=0}^{\lceil \frac{16}{221} n \rceil - 1} e^{0.4 - 0.9 \frac{j}{n-1}} + \sum_{j=\lceil \frac{16}{221} n \rceil}^{\lceil \frac{2}{23} n \rceil - 1} e^{2 - 23 \frac{j}{n-1}} = R(A)
 \end{aligned}$$

which establishes that B is the only stochastically stable convention.

Also in this case we can obtain a totally different prediction by ruling out one scenario. Consider the reduced game $G^{(-3)}$, obtained from G by deleting scenario G_3 , and let $T^{(-3)}$ be the associated transition matrix, obtained by deleting the third row and column of T and setting $T_{G_1, G_1}^{(-3)} = 0.992$. By applying again Proposition 5, we obtain that:

$$R(A) = \sum_{j=0}^{\lceil \frac{2}{23}n \rceil - 1} e^{2-23\frac{j}{n-1}} > \sum_{j=0}^{\lceil \frac{1}{11}n \rceil - 1} e^{1-11\frac{j}{n-1}} = R(B).$$

which establishes that A is the only stochastically stable convention in $G^{(-3)}$. This follows from the fact that the lowest expected payoff loss for making a mistake that leads to play A is attained in G_1 for $\phi_A \in [0, \frac{1}{11})$ and the lowest expected payoff loss for making a mistake that leads to play B is attained in G_2 for $\phi_B \in [0, \frac{2}{23})$.

A.2.2 Calculations of Example 1

Under uniform mistakes, all perturbations are equally likely. Hence, the pivotal scenario for convention selection is determined by the size of each convention's basin of attraction. Let $f_{\check{G}_k}^A$ denote the minimum fraction of the population that has to play action A in scenario $\check{G}_k$ to reach convention A with no mistakes in the process (i.e., it is a best response for all players). Define $f_{\check{G}_k}^B$ analogously. Then:

$$f_{\check{G}_k}^A = \frac{\pi_{\check{G}_k}^{BB} - \pi_{\check{G}_k}^{AB}}{\pi_{\check{G}_k}^{AA} + \pi_{\check{G}_k}^{BB} - \pi_{\check{G}_k}^{AB} - \pi_{\check{G}_k}^{BA}} ; \quad f_{\check{G}_k}^B = \frac{\pi_{\check{G}_k}^{AA} - \pi_{\check{G}_k}^{BA}}{\pi_{\check{G}_k}^{AA} + \pi_{\check{G}_k}^{BB} - \pi_{\check{G}_k}^{AB} - \pi_{\check{G}_k}^{BA}} .$$

Under uniform mistakes, to compute the radius of each convention is enough to take the complement to one of the basins of attraction of the other convention. See Table 16 for the relevant calculations.

Then, denoting as $n_k(Z) = \lceil f_{\check{G}_k}^Z n \rceil$ the minimum number of agents playing action $Z \in \{A, B\}$ needed in order to make Z a best response in

Scenario	For R(A)	For R(B)
$\check{G}_1$	$f_{\check{G}_1}^B = 10/11 \simeq 0.9091$	$f_{\check{G}_1}^A = 1/11 \simeq 0.0909$
$\check{G}_2$	$f_{\check{G}_2}^B = 2/23 \simeq 0.0870$	$f_{\check{G}_2}^A = 21/23 \simeq 0.9130$
$\check{G}_3$	$f_{\check{G}_3}^B = 4/9 \simeq 0.4444$	$f_{\check{G}_3}^A = 5/9 \simeq 0.5556$

Table 16: Scenario characteristics that are relevant under uniform mistakes. $f_{\check{G}_k}^Z$ is the minimum fraction of agents playing action $Z \in \{A, B\}$ needed to make Z a best response in scenario $\check{G}_k$.

scenario $\check{G}_k$ for all agents, we have that

$$\begin{aligned}
R(A) &= \min_k \{R_k(A)\} = \min_k \{n_k(B) r\} = \min_k \left\{ \left\lceil f_{\check{G}_k}^B n \right\rceil \right\} r \\
&= \left\lceil f_{\check{G}_2}^B n \right\rceil r = \left\lceil \frac{2}{23} n \right\rceil r \\
R(B) &= \min_k \{R_k(B)\} = \min_k \{n_k(A) r\} = \min_k \left\{ \left\lceil f_{\check{G}_k}^A n \right\rceil \right\} r \\
&= \left\lceil f_{\check{G}_1}^A n \right\rceil r = \left\lceil \frac{1}{11} n \right\rceil r
\end{aligned}$$

Therefore, $R(B) > R(A)$, so convention B is the unique stochastically stable convention in game $\check{G}$.

If we consider the reduced game $\check{G}^{(-2)}$, obtained by deleting scenario $\check{G}_2$, and let $\check{T}^{(-2)}$ be the associated transition matrix obtained by deleting the second row and column of $\check{T}$ and setting $\check{T}_{\check{G}_1, \check{G}_1}^{(-2)} = 0.995$, we obtain

$$\begin{aligned}
R(A) &= \min_k \{R_k(A)\} = \min_k \{n_k(B) r\} = \min_k \left\{ \left\lceil f_{\check{G}_k}^B n \right\rceil \right\} r \\
&= \left\lceil f_{\check{G}_3}^B n \right\rceil r = \left\lceil \frac{4}{9} n \right\rceil r \\
R(B) &= \min_k \{R_k(B)\} = \min_k \{n_k(A) r\} = \min_k \left\{ \left\lceil f_{\check{G}_k}^A n \right\rceil \right\} r \\
&= \left\lceil f_{\check{G}_1}^A n \right\rceil r = \left\lceil \frac{1}{11} n \right\rceil r
\end{aligned}$$

So $R(A) > R(B)$, and convention A turns out to be the unique stochastically stable convention in game $\check{G}^{(-2)}$.

A.2.3 Calculations of Example 2

Under the intentional mistakes model, agents make errors only if these can lead to a preferred equilibrium in the current scenario. Table 17 reports the payoff differences between convention A and convention B for each scenario of the game.

Scenario	For $R(A)$	For $R(B)$
$\dot{G}_1$	$\pi_{\dot{G}_1}^{BB} - \pi_{\dot{G}_1}^{AA} = -9$	$\pi_{\dot{G}_1}^{AA} - \pi_{\dot{G}_1}^{BB} = 9$
$\dot{G}_2$	$\pi_{\dot{G}_2}^{BB} - \pi_{\dot{G}_2}^{AA} = 1$	$\pi_{\dot{G}_2}^{AA} - \pi_{\dot{G}_2}^{BB} = -1$
$\dot{G}_3$	$\pi_{\dot{G}_3}^{BB} - \pi_{\dot{G}_3}^{AA} = 19$	$\pi_{\dot{G}_3}^{AA} - \pi_{\dot{G}_3}^{BB} = -19$

Table 17: Payoff differences between convention A and convention B in each scenario.

For simplicity, assume that the resistance to mistakes is equal to the inverse of the payoff difference between conventions when that difference is positive, and infinite otherwise.

Then, $\min_k \{r_k^B(\pi_{\dot{G}_k}^{BB} - \pi_{\dot{G}_k}^{AA})\} = \frac{1}{19} < \frac{1}{9} = \min_k \{r_k^A(\pi_{\dot{G}_k}^{AA} - \pi_{\dot{G}_k}^{BB})\}$. Besides, for n large enough, $\min_k \{n_k(B)\} = \lceil \frac{2}{23}n \rceil < \lceil \frac{1}{11}n \rceil = \min_k \{n_k(A)\}$.

Therefore, applying Corollary 1, we get that

$$R(A) = \min_k \{r_k^B(\pi_{\dot{G}_k}^{BB} - \pi_{\dot{G}_k}^{AA})\} \min_l \{n_l(B)\} = \frac{1}{19} \lceil \frac{2}{23}n \rceil$$

$$R(B) = \min_k \{r_k^A(\pi_{\dot{G}_k}^{AA} - \pi_{\dot{G}_k}^{BB})\} \min_l \{n_l(A)\} = \frac{1}{9} \lceil \frac{1}{11}n \rceil$$

Then, $R(B) > R(A)$, so convention B is the unique stochastically stable convention in game $\dot{G}$.

If we consider the reduced game $\dot{G}^{(-3)}$, obtained by deleting scenario $\dot{G}_3$, and let $\dot{T}^{(-3)}$ be the associated transition matrix obtained by deleting the third row and column of $\dot{T}$ and setting $\dot{T}_{\dot{G}_1, \dot{G}_1}^{(-3)} = 0.995$, we obtain

$$\min_{k \neq 3} \{r_k(\pi_{\dot{G}_k}^{AA} - \pi_{\dot{G}_k}^{BB})\} = \frac{1}{9} > \frac{1}{1} = \min_{k \neq 3} \{r_k(\pi_{\dot{G}_k}^{BB} - \pi_{\dot{G}_k}^{AA})\}.$$

Then, applying Corollary 1, we obtain that

$$R(A) = \min_{k \neq 3} \{r_k^B(\pi_{\dot{G}_k}^{BB} - \pi_{\dot{G}_k}^{AA})\} \min_{l \neq 3} \{n_l(B)\} = \frac{1}{1} \left\lceil \frac{2}{23}n \right\rceil$$

$$R(B) = \min_{k \neq 3} \{r_k^A(\pi_{\dot{G}_k}^{AA} - \pi_{\dot{G}_k}^{BB})\} \min_{l \neq 3} \{n_l(A)\} = \frac{1}{9} \left\lceil \frac{1}{11}n \right\rceil$$

So $R(A) > R(B)$, and convention A turns out to be the unique stochastically stable convention in game $\dot{G}^{(-3)}$.

Appendix B

Appendix to Chapter 3

B.1 Analysis of senders' behavior in the strategic game

To test the effect of the level of uncertainty about the receiver's behavior on the message sent in the strategic game, we conducted several Logit regressions, whose results are presented in Table 18. There, we predict the message in the strategic game as a (logistic) function of the message in the non-strategic game, the answer to the belief question S_1 , and the distance from 50% of the answer to the belief question S_2 .

Therefore, the functional form of the general regression is:

$$\text{Message}_{\text{Str}} = \text{Logit} [\beta_0 + \text{Message}_{\text{NS}} + S_1 + f(50 - S_2)] \quad (\text{B.1})$$

Where $\text{Message}_{\text{Str}}$ is a dummy variable that takes value 1 if the message sent in the strategic game is false and 0 otherwise, $\text{Message}_{\text{NS}}$ is an analog variable for the non-strategic game, S_1 is a dummy variable that takes value 1 if the sender's expected option chosen by the receiver in the strategic game is B, $\text{Logit}(\cdot)$ is the standard logistic function, and $f(50 - S_2)$ is a function that represents the uncertainty about the receiver's

	(1)	(2)	(3)
Message _{NS}	1.716*** (0.216)	1.707*** (0.215)	1.713*** (0.214)
S ₁	1.966*** (0.209)	1.949*** (0.209)	1.974*** (0.210)
$f(50-S_2)$	0.0006 (0.007)	$7.469 \cdot 10^{-5}$ $(1.694 \cdot 10^{-4})$	-0.004 (0.005)
Intercept	-2.015*** (0.227)	-2.041*** (0.207)	-2.054*** (0.201)
N	593	593	593

Notes: Coefficient estimates from logit regressions where the dependent variable is lying in the strategic game, which takes the value of 1 if the sender sends the false message and 0 otherwise. Standard errors are presented in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 18: Effect of uncertainty in the message sent in the strategic sender-receiver game.

behavior in the strategic game. Column (1) shows the regression estimates of Model B.1 with $f(50-S_2) = |50 - S_2|$, Column (2) shows the results with $f(50-S_2) = (50 - S_2)^2$, and Column (3) shows the results with $f(50-S_2) = (50 - S_2)$.

As we can observe in Table 18, senders' uncertainty about the receiver's behavior in the strategic game has no statistically significant effect on the message sent by the senders after controlling for the message sent in the non-strategic game, and the answer to belief question S₁.

B.2 Comparison of the results in our data with the literature

Since we replicated the design in Sutter (2009) in our strategic sender-receiver game using the same belief elicitation questionnaire, we can compare the results from the data in this game in both samples. Tables 19 and 20 present the frequencies of false messages in our strategic game and whether senders expected to be followed in our online subjects' data and those in Sutter (2009). As it can be observed in both tables, we obtain similar results to the ones in the literature¹ (Gneezy, 2005; Sutter, 2009) regarding the sender's action (i.e., the message sent) and the sender's expectation to be followed in the strategic game for each payment treatment. This is confirmed through the six Pearson's Chi-squared tests we conducted, presented in the last row of each corresponding column.

False message	T ₁	T ₂	T ₃
Sutter (2009)	42/96 (44%)	34/96 (35%)	55/93 (59%)
Our study	102/202 (50.5%)	63/180 (35.0%)	127/211 (60.2%)
χ^2 test <i>p</i> -value	0.335	1	0.964

Table 19: Comparison of false messages sent in the strategic task by payment treatment in Sutter (2009) with our online subjects data. The last row displays the *p*-values of the Pearson's χ^2 tests that compare the frequency in both samples.

As it can be observed in the last row of Tables 19 and 20, there is no statistically significant difference at 5% level for all comparisons but in the expectation to be followed in payment treatment T₂ (Pearson's Chi-squared test *p*-value = 0.079), where both samples have a different result (more senders expected to be followed in our online data).

Finally, we can use the previous method in the literature to identify

¹The results regarding the frequency of each message sent obtained in Gneezy (2005) and Sutter (2009) are similar. Hence, for the sake of readability, we just report the results from Sutter (2009), who also had information regarding the senders' beliefs which can be compared to our data.

Expect trust	T ₁	T ₂	T ₃
Sutter (2009)	70/96 (73%)	67/96 (70%)	61/93 (66%)
Our study	156/202 (77.2%)	144/180 (80.0%)	156/211 (73.9%)
χ^2 test <i>p</i> -value	0.504	0.079	0.179

Table 20: Comparison of the number of senders who expected the receiver to follow the message in the strategic task by payment treatment in Sutter (2009) with our online data.

deceptive intentions, combining the sender's message in the strategic game with the expected option chosen by the receiver. Table 21 presents the frequencies for each of the four behaviors identified with the previous method in the literature in our data and in Sutter (2009), controlling by payment treatment.

Sender's behavioral type using the previous method in the literature	Sutter (2009) data			Our online data results		
	T ₁	T ₂	T ₃	T ₁	T ₂	T ₃
	N=96	N=96	N=93	N=202	N=180	N=211
Benevolent truth-teller	31%	36%	9%	36.1%	55.6%	28.4%
Sophisticated truth-teller	25%	28%	32%	13.4%	9.4%	11.4%
Plain liar	41%	33%	56%	41.1%	24.4%	45.5%
Benevolent liar	2%	2%	3%	9.4%	10.6%	14.7%

Table 21: Frequencies of the behavioral types obtained using the previous method in the literature (defined in Sutter (2009)) for the data in (Sutter, 2009) compared to our results with online subjects.

As it can be observed by comparing the frequencies between the two samples, we obtain significantly different frequencies in our data from those in Sutter (2009) for each behavior identified using the previous method in the literature, both controlling for each payment treatment and pooling them (Fisher's exact test *p*-values is lower than 0.001 aggregating payment treatments, 0.014 for T₁, and lower than 0.001 for T₂ and T₃). Hence, we cannot replicate the results of the deceptive intentions captured with the previous method in the literature obtained by Sutter (2009) using

our data of the strategic sender-receiver game.

In particular, compared to the laboratory results in Sutter (2009), we obtained a lower number of sophisticated truth-tellers but a higher number of benevolent liars using the previous method in the literature in our data. A possible explanation for these differences is that, whereas the laboratory setting promotes more deliberation, online subjects are faster in their cognitive reasoning, using simpler deceptive methods.

Moreover, our online sample subjects seem to care more about the receiver's payment than participants from previous laboratory studies since we obtained fewer senders with a deceptive intention than Sutter's laboratory results in the payment treatments where deceiving is more economically harmful. One reason that could explain the difference in this result is the distinct sample provenance, being our data from an online platform with US participants collected in 2021 in contrast to the laboratory sample from a German pool collected in 2005 by Sutter (2009).

B.3 Robustness analyses ruling out uncertain and incoherent senders

To test the robustness of our main findings, we replicate our analyses after excluding those senders who showed an inconsistency between their answers in the two belief questions and the action taken in the strategic sender-receiver game. In particular, a sender is denoted as *incoherent* if the answer to belief question S_2 (i.e., the probability of the message being trusted) does not align with the answer to belief question S_1 (i.e., the expected action taken by the matched receiver) given the message sent in the strategic sender-receiver game. This could happen either if $S_2 > 50$ but the expected option chosen by the receiver is different from the option suggested in the message, or if $S_2 < 50$ but the expected option chosen by the receiver is equal to the option suggested in the message.

As it can be observed by comparing the first two columns in Table 22,

there is no statistically significant difference in the frequencies of each behavioral type obtained with our new method after excluding incoherent senders (Pearson’s Chi-squared test p -value = 0.61). This applies to both the absolute frequency of senders included in each category detected with our new method and the relative frequencies from the categories obtained using the previous approach in the literature.

Moreover, we found that 15.85% of the senders in our sample provided incoherent answers to questions S_1 and S_2 . This proportion is not statistically different from the one obtained in the literature (Sutter (2009) got 12.78%, which is not significantly different, Pearson’s Chi-squared test p -value = 0.27). Finally, the rest of results in the paper remain unchanged (sometimes having a stronger statistical significance) after ruling out incoherent senders.

Sender’s behavioral type with our method	(1) (N=593)	(2) (N=499)	(3) (N=477)	(4) (N=383)
Sophisticated truth-teller	60%	64%	61%	67%
Pessimistic truth-teller	40%	36%	39%	33%
Honest truth-teller	76%	77%	76%	77%
Incongruous truth-teller	24%	23%	24%	23%
Benevolent liar	26%	24%	18%	5%
Doubly liar	74%	76%	82%	95%
Plain liar	84%	86%	87%	90%
Incongruous liar	16%	14%	13%	10%

Table 22: Comparison of sender classification under our new identification method across alternative sample restrictions. Column (1) uses the entire sample, Column (2) excludes senders with incoherent belief answers, Column (3) excludes uncertain senders (i.e., those with $S_2 = 50$), and Column (4) excludes both incoherent and uncertain senders.

Complementing our results about the lack of significance of the answers to question S_2 given the message sent in the strategic game and

the answers to question S_1 , we analyzed how our main results were influenced by senders who expressed complete uncertainty, i.e., $S_2 = 50$. The exclusion of these senders resulted in a final sample of 477 subjects. The relative frequencies of the categories from our new method are presented in the third column of Table 22. As observed, there is no statistically significant difference in the frequencies obtained for each behavioral type categorized with our new method after ruling out fully uncertain senders (Pearson’s Chi-squared test p-value = 0.92).

To be more restrictive, we analyzed the data after excluding both incoherent and fully uncertain senders, ending up with 383 participants. The last column in Table 22 presents the relative frequency of the behavioral types obtained with our new method. The only difference with our entire sample is that there is a significantly higher proportion of doubly liars (Fisher’s exact test p-value = 0.035) and a marginally lower proportion of incongruous liars (Fisher’s exact test p-value = 0.097). Notice that, if any, these two results could just strengthen our main conclusions and the take-home message of our paper.

Finally, to provide more evidence on how uncertainty could affect our main findings, we conducted a similar analysis discarding senders who are incoherent and those whose $S_2 \in [40, 60]$, ending up with 313 participants. The comparison with the entire sample and after discarding incoherent and fully uncertain senders is presented in Table 23. As it can be observed by comparing the last two columns in the table, there is no significant difference between the results discarding fully uncertain or with $S_2 \in [40, 60]$ with no incoherent senders (Pearson’s Chi-squared test p-value = 0.215). This confirms the robustness of our main results.

B.4 Further analyses to check if random behavior drives our results

To strengthen our results against the hypothesis of randomness fully driving them, we computed the same tests for random behavior considering

Sender's behavioral type with our method	(1) (N=593)	(2) (N=313)	(3) (N=383)
Sophisticated truth-teller	60%	71%	67%
Pessimistic truth-teller	40%	29%	33%
Honest truth-teller	76%	75%	77%
Incongruous truth-teller	24%	25%	23%
Benevolent liar	26%	7%	5%
Doubly liar	74%	93%	95%
Plain liar	84%	93%	90%
Incongruous liar	16%	7%	10%

Table 23: Comparison of sender classification under our new identification method across alternative sample restrictions. Column (1) uses the entire sample, Column (2) excludes senders with incoherent belief answers and those with $S_2 \in [40, 60]$, Column (3) excludes incoherent senders and those uncertain (i.e., with $S_2 = 50$).

all non-plain categories (i.e., all classes but honest truth-tellers and plain liars) instead of just the newly defined ones presented in Section 3.3.2.

As in the newly defined categories case, the six non-plain categories are jointly different from their respective average, which is 6.44% (Proportion test p-value < 0.001). Besides, we can reject the null hypothesis that the behavior of a sender in our data can be categorized with the same probability in each of the non-plain behavioral types (Fisher's exact test p-value < 0.001), which would be another potential indication of random behavior driving our main findings.

Furthermore, we checked for the possibility that a fraction of senders playing randomly drives the observed results. To do so, we conducted two series of tests: First, we carried out proportion tests against the null hypothesis that $x\%$ of the senders in the data are spread uniformly over the four new behavioral types (as it would be if these are driven by random behavior). We consider each noise level from 1% to 100%. Then, considering the same percentage levels of randomness, we conducted an

individual binomial test for each behavioral type categorized with our new method in order to check if this level of randomness could explain the fraction of data we obtained. The results of these batteries of tests are:

- There is no level of randomness to which the frequency of the four new behavioral types is jointly equal, so we can reject the hypothesis that the four new behavioral types are altogether totally driven by a fraction of players playing randomly.
- For pessimistic truth-tellers, we cannot reject the hypothesis that $x\%$ of the senders playing randomly explains the number of senders categorized in this behavioral type if the fraction of senders x is between 25% (3.125% for each class) and 52% (6.5% for each category) of the whole number of senders in our data.
- For incongruous liars, we cannot reject this hypothesis for total randomness levels between 35% (4.375% for each class) and 66% (8.25% for each category).
- For incongruous truth-tellers, we cannot reject this hypothesis for total randomness levels between 59% (7.375% for each class) and 96% (12% for each category).
- For doubly liars, we cannot reject this hypothesis for total levels of randomness between 53% (6.625% for each class) and 89% (11.125% for each category).

Notice that the levels for which we cannot reject the hypothesis for each new behavioral type are mutually exclusive, that is, the intersection of the four sets is empty. Hence, there is no percentage of senders playing randomly that can explain the data we obtained in the four new behavioral types at the same time.

B.5 Main results controlling by order condition

As presented in Section 3.3.1, Tables 24 and 25 present the frequency of false messages in the strategic sender-receiver game and whether senders expect to be followed in the two order conditions.

As shown in Table 24, and confirmed using Fisher exact tests, there are no order effects in the message sent in the strategic sender-receiver game. Fisher exact test p-values are 0.218 aggregating payment treatments, 0.328 for T_1 , 0.640 for T_2 , 0.675 for T_3 .

False message	T_1	T_2	T_3	Total
Normal order	50/106 (47.2%)	33/99 (33.3%)	64/109 (58.7%)	147/314 (46.8%)
Inverted order	52/96 (54.2%)	30/81 (37.0%)	63/102 (61.8%)	145/279 (52.0%)

Table 24: Comparison of false messages sent in the strategic task by payment treatment in the two order conditions.

In Table 25, we observe that the expectation to be followed in the strategic sender-receiver game is similar for the two order conditions. Using Fisher exact tests, we confirm that there are no statistical differences between the two order conditions: p-values are 0.558 aggregating payment treatments, 0.615 for T_1 , 0.264 for T_2 , and 0.754 for T_3 .

Expect trust	T_1	T_2	T_3	Total
Normal order	80/106 (75.5%)	76/99 (76.8%)	82/109 (75.2%)	238/314 (75.8%)
Inverted order	76/96 (79.2%)	68/81 (84.0%)	74/102 (72.5%)	218/279 (78.1%)

Table 25: Comparison of the number of senders who expect the receiver to follow the message of the strategic task by payment treatment in the two order conditions.

Table 26 displays the frequencies obtained for all the behavioral types categorized with our new method controlling by the order in which the two games are presented. In this table, we observe the lack of order effects in our data, as shown in Section 3.3.2. Inspecting each behavioral type individually, the only category that leads to differences depending on

the order in which the games are presented is the incongruous liar one (Fisher's exact test p-value = 0.025).

Sender's behavioral type with our new method	Normal order	Inverted order
Sophisticated truth-teller	26 (8.3%)	15 (5.4%)
Pessimistic truth-teller	14 (4.5%)	13 (4.7%)
Honest truth-teller	94 (29.9%)	83 (29.7%)
Incongruous truth-teller	33 (10.5%)	23 (8.2%)
Benevolent liar	8 (2.5%)	10 (3.6%)
Doubly liar	28 (8.9%)	23 (8.2%)
Plain liar	99 (31.5%)	88 (31.5%)
Incongruous liar	12 (3.8%)	24 (8.6%)

Table 26: Frequencies of the behavioral types obtained with our new method controlled by order.

Figures 18 and 19 show the decomposition of senders in each category of the previous method on the behavioral types captured with our new method for each order condition.

B.6 Experimental procedure diagrams

The diagrams in Figure 20 represent the procedure followed by senders who played in the inverted order condition. Figures 21 and 22 represent the procedure followed by receivers in both order conditions. These procedures are explained in further detail in Section 3.2.3.

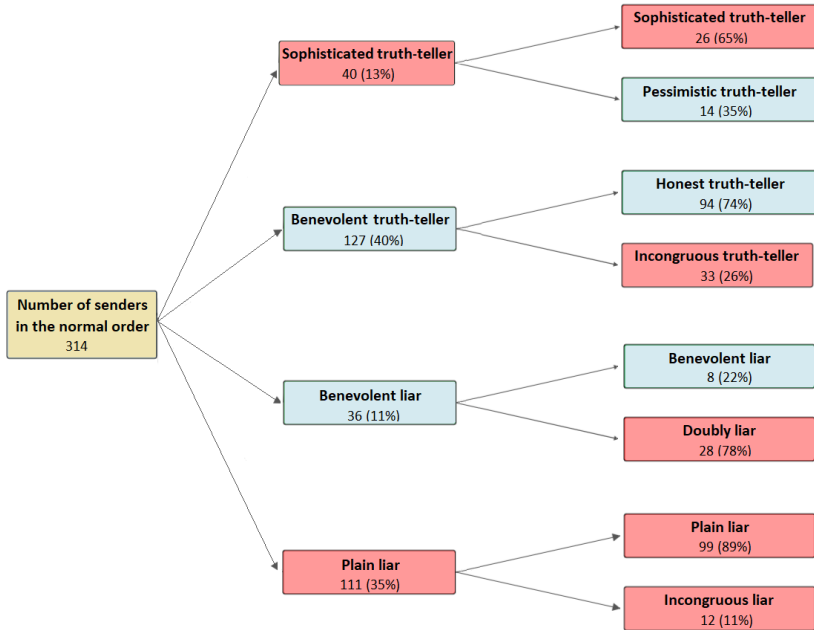


Figure 18: Decomposition of the behavioral types categorized with the previous method into the new behaviors categorized with our new method for the normal order condition. Percentages in parentheses are the relative frequencies with respect to the respective parent node.

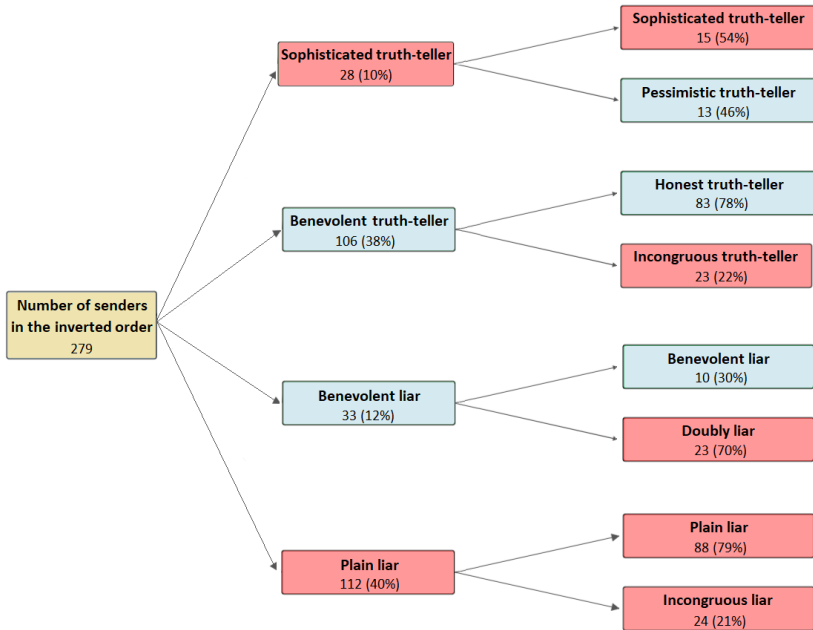


Figure 19: Decomposition of the behavioral types categorized with the previous method into the new behaviors categorized with our new method for the inverted order condition. Percentages in parentheses are the relative frequencies with respect to the respective parent node.

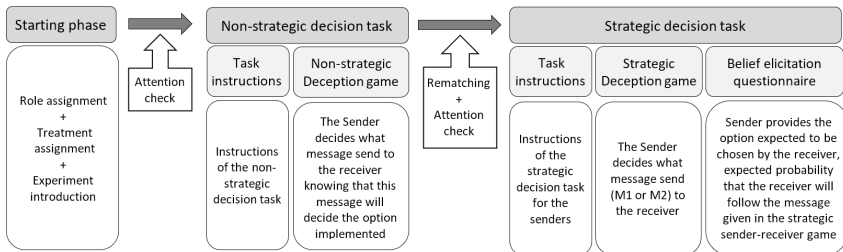


Figure 20: Procedure of the senders who play in the inverted order condition.

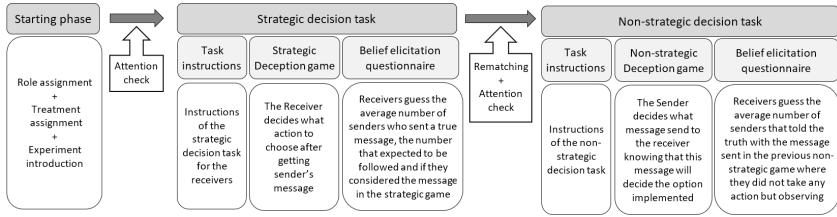


Figure 21: Procedure of the receivers who play in the normal order condition.

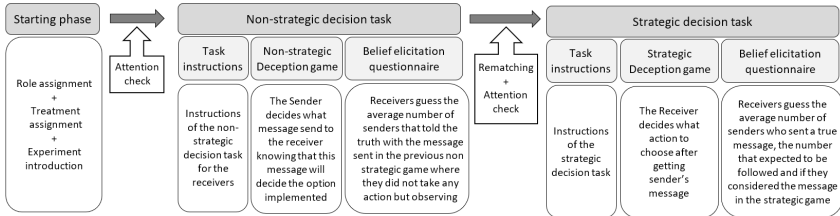


Figure 22: Procedure of the receivers who play in the inverted order condition.

B.7 Sociodemographic characteristics

Table 27 displays descriptive information about the sociodemographic characteristics of the participants in our subject pool.

We found no significant effect of age in the fraction of senders identified with a deceptive intention. This result holds either using our new method (t-test p-value = 0.097), or for those identified using the previous approach in the literature (p-value = 0.16), or identifying senders only using the veracity of their message in the strategic game (p-value = 0.07).

Moreover, we conducted additional analyses separating older participants (those with an age higher than the average of the whole sample) from younger subjects. The average of the whole sender pool is 30.12 years old. Although it seems that younger participants might deceive more than the older in our subject pool (65.5% vs. 59.2%), we obtained no statistically significant difference between the two groups in the fraction of subjects identified with a deceptive intention using our new method

Demographic	Category	Senders		Receivers	
		N	%	N	%
Gender	Female	298	50.3%	296	49.9%
	Male	295	49.7%	297	50.1%
Religion	Non-religious	311	52.4%	280	47.2%
	Christianity	191	32.2%	214	36.1%
	Islam	12	2.0%	11	1.9%
	Spiritualism	14	2.4%	16	2.7%
	Other	65	11.0%	72	12.1%
Ethnicity	White/Caucasian	382	64.4%	413	69.6%
	Latino/Hispanic	43	7.3%	45	7.6%
	Mixed	42	7.1%	29	4.9%
	Black	34	6.2%	41	6.9%
	East Asian	42	7.1%	26	4.4%
	Other	47	7.9%	39	6.6%
Student status	No	409	69.0%	449	75.7%
	Yes	166	28.0%	136	22.9%
	Other	18	3.1%	8	1.3%
Employment status	Full-time	213	35.9%	286	48.2%
	Unemployed	119	20.1%	91	15.3%
	Part-time	110	18.5%	92	15.5%
	Not in paid work	66	11.1%	54	9.1%
	Other	85	14.3%	70	11.8%
Country of birth	United States	561	94.6%	558	94.1%
	Other	32	5.4%	35	5.9%
First language	English	545	91.9%	549	92.6%
	Spanish	14	2.4%	13	2.2%
	Chinese	9	1.5%	9	1.5%
	Other	25	4.2%	22	3.7%

Table 27: Sociodemographic characteristics of the sender and receiver participant pools.

(Pearson's Chi-squared test p-value = 0.18). This result holds using the previous method in the literature to identify deceptive intentions (51.3% for younger subjects vs. 45.97% for older subjects, Pearson's Chi-squared test p-value = 0.23). All results are preserved if we use the median (which is 29 years old) instead of the mean to separate younger (with an age lower or equal to the median) from older participants.

Furthermore, we conducted analogous tests to study gender differences in the deceptive intention of our subjects. Although we observe a trend suggesting that females aim to deceive more than men in our subject pool, we obtain no statistically significant difference between the proportion of men and female with a deceptive intention neither using our new method (Pearson's Chi-squared test p-value = 0.06) nor the previous method in the literature (Pearson's Chi-squared test p-value = 0.48).

B.8 Instructions

We present screenshots of the instructions provided in the experiment for participants playing as senders and receivers. For the sake of exposition and to avoid repetition, we just display the instructions presented to those participants who played in payment treatment T_1 in the normal order condition. Instructions for treatments T_2 and T_3 are identical, varying the payoff numbers accordingly. The few differences between the instructions provided in both orders are explained in the caption of each correspondent page. These changes were made to be coherent throughout the whole experiment, e.g., referring to a task as first or second.

End of the study

Unfortunately, you did not pass the attention check in the previous page, so you cannot be paid for this study.

As we explained in the study description, you will not be able to continue, not obtaining the final completion code.

Please, close the tab or go to the Prolific webpage by clicking the following button.

[Go to Prolific webpage](#)

Figure 23: Page presented to those senders who failed an attention check.

Figure 23 presents the page displayed to those senders who failed any of the two attention checks conducted, which are placed in the experiment introduction (Figure 24 for senders and Figure 29 for receivers) and on the page between both games, where participants are informed they are rematched (Figure 27 for senders and Figure 35 for receivers).

Welcome to this study!

General information:

During this study, you will make some decisions in different tasks that allow you to earn some extra money.

Your decisions and earnings are confidential: we will not inform any other participant about the money you earned, nor we will identify you with your choices during the session. Your earnings will depend on your decisions as well as on the decisions made by other participants.

The session:

You will face a sequence of tasks that can involve another randomly picked participant. **In each task, you will be matched with a different and new participant.**

You are now **randomly assigned a role** between Person 1 and Person 2, which you can see in the box below. Your role will affect which decisions you can make and what payments you can get. Remember: **You will keep the same role in all tasks.**

Your role:

Person 1

Payments:

You will get an extra **payment determined according to the decisions made in a single task.**

Such single task will be **randomly selected at the end of this study.**

Your payment will be determined by the **points obtained in the selected task.** Also, the participant in the role of **Person 2 in the selected task will be paid according to the points obtained in that task.**

This line checks if you are reading carefully the page: after reading the whole page, press the number 1 key in your keyboard (either above the letter keys or enabling your numeric keypad) instead of clicking the next button at the end of the page in order to continue in the study (if you click the next button you will not go to the next page, not receiving any completion code and then no payment will be received). Notice that this is not necessarily the last attention check you will have to pass along this study.

You are the only participant who will be ever informed about the points you earn in all tasks.

The **conversion rate** is: 10 points = 0.5 GBP. You will earn between 5 and 15 points in this study.

After carefully reading all the previous instructions, you can proceed with the study.

Next

Figure 24: Experiment introduction page presented to senders.

B.8.1 Senders' instructions

Participants in the normal order condition were shown the figures in the order presented below, that is, Figure 24-25-26-27-28. Participants who played in the inverted order condition were shown the analogous figures in the following order: Figure 24-28-27-25-26.

First task decision: send one message.

The first task

Remember: your role is Person 1.

There are two options (A and B) that can be chosen in this task.
Both options yield different payoffs for you (Person 1) and the Person 2 that is paired with you.

The options are:

- Option A: 5 points for you (Person 1) and 6 points for Person 2.
- Option B: 6 points for you (Person 1) and 5 points for Person 2.

Person 2 will decide which option to implement. Yet, Person 2 will never be informed about the points associated with either option. Instead, before making a decision, Person 2 will only receive one of the two following messages:

- Message 1: "Option A will earn you more money than Option B".
- Message 2: "Option B will earn you more money than Option A".

You (Person 1) have to choose which message you want to send to Person 2.

To repeat, Person 2 will never be informed about your points in the chosen option, nor about the possible points (for both persons) in the option not chosen.

What message do you send?

- ☐ Option A will earn you more money than Option B
- ☐ Option B will earn you more money than Option A

Press the next button after choosing the message you want to send to Person 2.

Next

Figure 25: Senders' instructions in the strategic task for the normal order condition. The instructions for those who played in the inverted order condition are similar, replacing the word "first" with "second".

Please, answer the following question about this first task.

Your answers in this form will be **private**, so no other participant will ever know it.

Which option do you expect Person 2 to choose?

☐ Option A ☐ Option B

Out of 100 participants with Person 2 role, how many do you think follow Person 1 message on average? Write your answer between 0 and 100:

Please, press the following button after answering the question above.

Next

Figure 26: Senders' page where the belief questions were elicited for the normal order condition. The instructions for those who played in the inverted order condition are similar, replacing the word "first" with "second", which is written in bold letters and underlined.

We remind you that you are going to be **matched with a new participant for the next task.**

No other participant will ever be informed about the message you sent in the previous task.

No participant with the role of Person 2 has yet been informed of the points earned in case the previous task is selected.

Your identity and answers in all the forms provided during this study are **private**, so none of your matched Person 2 will know them neither during nor after the whole study.

Remember that you will be **paid according to the outcome of one randomly chosen task** of the whole study.

This task will be **randomly determined after you make all your decisions.**

This line checks if you are reading carefully the page: after reading the whole page, press the number 9 key in your keyboard (either above the letter keys or enabling your numeric keypad) instead of clicking the next button at the end of the page in order to continue in the study (if you click the next button you will not go to the next page, not receiving any completion code and then no payment will be received). Notice that this is not necessarily the last attention check you will have to pass along this study.

After carefully reading all the page, you can proceed with the study.

Next

Figure 27: Instructions shown to senders between the two games, stating that they are rematched with a different receiver.

Second task decision: send one message.

The second task

Remember: your role is **Person 1**.

There are two options (A and B) in this task.
Both options yield different payoffs for you (Person 1) and the Person 2 that is paired with you.

The options are:

- **Option A:** 5 points for you (Person 1) and 6 points for Person 2.
- **Option B:** 6 points for you (Person 1) and 5 points for Person 2.

Person 2 will never be informed about the points associated with either option.
Person 2 will only receive one of the two following messages:

- **Message 1:** "Option A will earn you more money than Option B".
- **Message 2:** "Option B will earn you more money than Option A".

You (Person 1) have to **choose which message you want to send** to Person 2.

Person 2 will make **NO** choice in this task. Instead, the message you send to Person 2 will determine the option selected, and then the payments of both of you. How the messages determine the option selected is:

- If you send **Message 1** ("Option A will earn you more money than Option B"), then **Option A is implemented**
- If you send **Message 2** ("Option B will earn you more money than Option A"), then **Option B is implemented**.

To repeat, Person 2 will **never** be informed about your points in the chosen option, nor about the possible points (for both persons) in the option not chosen.

What message do you send?

☐ Option A will earn you more money than Option B

☐ Option B will earn you more money than Option A

Press the next button after choosing the message you want to send to Person 2.

Next

Figure 28: Senders' instructions page in the non-strategic task for the normal order condition. The instructions for those who played in the inverted order condition are similar, replacing the word "second" with "first".

B.8.2 Receivers' instructions

Receivers in the normal order condition were shown the figures in the presented order, i.e., Figure 29-30-31-32-33-34-35-36-37-38. Those who played in the inverted order condition were shown analogous pages in the following order: Figure 29-36-37-38-35-30-31-32-33-34.

Welcome to this study!

General information:

During this study, you will make some decisions in different tasks that allow you to earn some extra money.

Your decisions and earnings are confidential: we will not inform any other participant about the money you earned, nor we will identify you with your choices during the session. Your earnings will depend on your decisions as well as on the decisions made by other participants.

The session:

You will face a sequence of tasks that can involve another randomly picked participant. **In each task, you will be matched with a different and new participant.**

You are now **randomly assigned a role** between Person 1 and Person 2, which you can see in the box below. Your role will affect which decisions you can make and what payments you can get. Remember: **You will keep the same role in all tasks.**

Your role:

Person 2

Payments:

You will get an extra payment determined according to the decisions made in a single task. Such single task will be randomly selected at the end of this study.

Your payment will be determined by the points obtained in the selected task. Also, the participant in the role of Person 1 in the selected task will be paid according to the points obtained in that task.

This line checks if you are reading carefully the page: after reading the whole page, press the number 3 key in your keyboard (either above the letter keys or enabling your numeric keypad) instead of clicking the next button at the end of the page in order to continue in the study (if you click the next button you will not go to the next page, not receiving any completion code and then no payment will be received). Notice that this is not necessarily the last attention check you will have to pass along this study.

You are the only participant who will be ever informed about the points you earn in all tasks.

The conversion rate is: 10 points = 0.5 GBP. You will earn between 5 and 15 points in this study.

After carefully reading all the previous instructions, you can proceed with the study.

Next

Figure 29: Experiment introduction page presented to receivers.

First task instructions.

The first task

Remember: your role is **Person 2**.

Two possible payoffs are available to you and the Person 1 paired with you in this task. These payoffs depend on the option (A or B) chosen by **you** (Person 2).

We showed all payoffs (in points) associated with each option to Person 1. **In this task, the only information you will have before making your choice is the message Person 1 sent to you.** Person 1 can choose one of the two following messages:

- **Message 1:** "Option A will earn you more money than Option B".
- **Message 2:** "Option B will earn you more money than Option A".

You will be informed about the message chosen by Person 1 on the next page.

We will then ask you to **choose either Option A or Option B**.

Your choice will determine the payoffs of both you and Person 1 in this task.

You will **never** know what points were offered in the option not chosen, nor the points that Person 1 could earn in any option.

After reading the instructions, press the next button to receive Person 1's message and choose one option.

Next

Figure 30: Receivers' instructions page in the strategic task for the normal order condition. The instructions for those who played in the inverted order condition are similar, replacing the word "first" with "second".

First task decision: choose one option.

Person 1 sent you:

Message 1: "Option A will earn you more money than Option B".

What option do you choose?

☐ Option A ☐ Option B

Press the next button after choosing the option to implement in this task.

Next

Figure 31: Receivers' decision page in the strategic task for the normal order condition. The instructions for those who played in the inverted order condition are similar, replacing the word "first" with "second".

Please, answer the following question about this first task.

Your answer in this form will be **private**, so no other participant will ever know it.

Out of 100 participants with Person 1 role, how many do you think sent a true message? Write your answer between 0 (= all sent a false message) and 100 (= all sent a true message):

Please, press the following button after answering the question above.

Next

Figure 32: Receivers' page where the belief question R_1 in the strategic task was elicited for the normal order condition. The instructions for those who played in the inverted order condition are similar, replacing the word "first" with "second", which is written in bold letters and underlined.

Your answer in this form will be **private**, so no other participant will ever know it.

Out of 100 participants with Person 1 role, how many do you think that expect Person 2 to follow the message? Write your answer between 0 and 100:

Please, press the following button after answering the question above.

Next

Figure 33: Receivers' page where the belief question R_2 in the strategic task was elicited for the normal order condition. The same page is presented for those who played in the inverted order condition, adding at the beginning the first sentence in Figure 32 for the inverted order.

Your answer in this form will be **private**, so no other participant will ever know it.

Did you take into account the message received by Person 1 when you made your decision? Take into account encompasses both following the option recommended in Person 1's message and inverting it to the other option.

☐ Yes ☐ No

Please, press the following button after answering the question above.

Next

Figure 34: Receivers' page where the belief question R_3 in the strategic task was elicited for the normal order condition. The same page is presented for those who played in the inverted order condition, adding at the beginning the first sentence in Figure 32 for the inverted order.

We remind you that you are going to be **matched with a new participant for the next task**.
No other participant will ever be informed about the message you received in the previous task.
No participant with the role of Person 1 has yet been informed of the decision of participants with the role of Person 2 in the previous task.

Your identity and answers in all the forms provided during this study are **private**, so none of your matched Person 1 will know them neither during nor after the whole study.

Remember that you will be **paid according to the outcome of one randomly chosen task** of the whole study.
This task will be **randomly determined after you make all your decisions**. This line checks if you are reading carefully the page: after reading the whole page, press the number 7 key in your keyboard (either above the letter keys or enabling your numeric keypad) instead of clicking the next button at the end of the page in order to continue in the study (if you click the next button you will not go to the next page, not receiving any completion code and then no payment will be received). Notice that this is not necessarily the last attention check you will have to pass along this study.

After carefully reading all the page, you can proceed with the study.

Next

Figure 35: Instructions shown to receivers between the two games, stating that they are rematched with a different sender. The same page is presented for those who played in the inverted order condition but deleting the sentence “No participant with the role of Person 1 has yet been informed of the decision of participants with the role of Person 2 in the previous task.”.

Second task instructions.

The second task

Remember: your role is **Person 2**.

Two possible payoffs are available to you and the Person 1 paired with you in this task.
We showed all payoffs (in points) associated with each option to Person 1.

In this task, only Person 1 makes a decision: send you a message about both options. **This message will determine the option implemented** and then both persons' payoffs (in points). Person 1 can choose one of the two following messages:

- **Message 1:** "Option A will earn you more money than Option B".
- **Message 2:** "Option B will earn you more money than Option A".

You will be informed about the message chosen by Person 1 on the next page.

In this task, you will make NO choice. You will only know Person 1's message and the option implemented by this message.

You will **never** know what points were offered in the option not chosen, nor the points that Person 1 could earn in any option.

After reading the instructions of this task, press the next button to receive Person 1's message.

Next

Figure 36: Receivers' instructions page in the non-strategic task for the normal order condition. The instructions for those who played in the inverted order condition are similar but replacing word “second” by “first”.

Second task: Message received.

Person 1 sent you:

Message 1: "Option A will earn you more money than Option B".

You do not take any action on this task.

Based on the message sent by Person 1 to you, **Option A is implemented in this task.**

Press the next button to continue.

Next

Figure 37: Receivers' decision page in the non-strategic task for the normal order condition. The instructions for those who played in the inverted order condition are similar, replacing word "second" with "first".

Please, answer the following **question about this second task.**

Your answer in this form will be **private**, so no other participant will ever know it.

Out of 100 participants with Person 1 role, how many do you think sent a true message? Write your answer between 0 (= all sent a false message) and 100 (= all sent a true message):

Please, press the following button after answering the question above.

Next

Figure 38: Receivers' page where the belief question R_1 in the non-strategic task was elicited for the normal order condition. The instructions for those who played in the inverted order condition are similar, replacing the word "second" with "first", which is not written in bold letters nor underlined.

Appendix C

Appendix to Chapter 4

C.1 Supplementary results of punishment and norms

C.1.1 Results of messages that do not benefit the sender

This section provides the results of punishment, personal norms, and injunctive social norms for messages that do not benefit the sender, which imply that the sender and the receiver earn 10 points. Figure 39 shows the punishment descriptive comparisons for these messages.¹

For choices that do not benefit the sender in the harmful and neutral situation, punishment is practically negligible. Conversely, in line with the effect of lying in the absence of harm shown in Subsection 4.3.1, lying that does not benefit the sender (i.e., false reports in the Truth Game) increases punishment ($p < 0.001$ for all situations). However, even when the receiver is also economically worse off (Beneficial situation in the Truth Game), the increase in punishment driven by lying *per se* is small in absolute magnitude (0.88 points).

¹Note that the y-axis scale in Figure 39 ranges from 0 to 4, so it is smaller than the scale used in Figure 9, which ranges from 0 to 8.

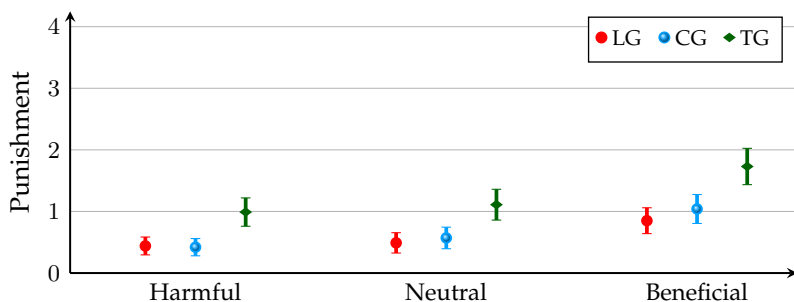


Figure 39: Punishment across games for messages that do not benefit the sender (i.e., truth-telling in the LG, lying in the TG and choosing the 10–10 split in the CG) by payoff situation.

These results are confirmed when examining the entire punishment distributions. As shown in Figure 40, punishment for lies that do not benefit the sender dominates that of their payoff-equivalent choices.

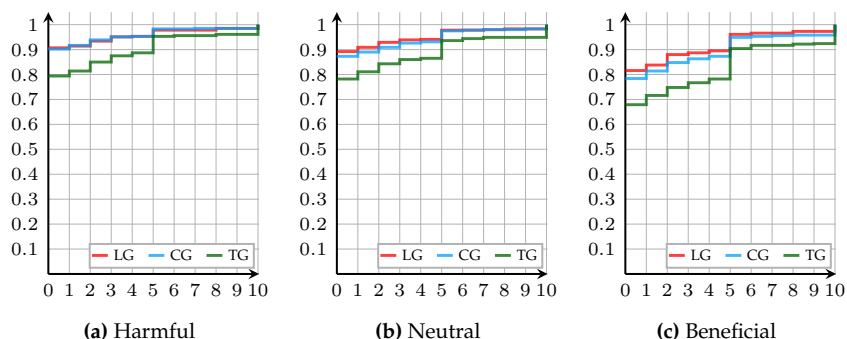


Figure 40: Empirical cumulative distribution functions of punishment for messages that do not benefit the sender by game and situation.

Around 80% of observers do not punish these selfless lies (Harmful and Neutral situations), whereas around 70% of observers do not punish a harmful false report if it harms the sender to the same extent. This latter result reinforces our interpretation that distributional concerns (especially those related to selfish inequality) are the main driver of punishment, rather than concerns about the message’s veracity.

On the other hand, punishment for truthful reports that do not benefit the sender dominates that of payoff-equivalent choices in the neutral and beneficial situations. However, since punishment for these messages is practically negligible, none of these differences are statistically significant (all Stuart–Maxwell test $p > 0.514$).²

We next present the normative assessments for messages that do not benefit the sender. Figures 41a and 41b show the descriptive results of personal and injunctive social norms, whereas Figures 42 and 43 show the empirical cumulative distribution functions.³

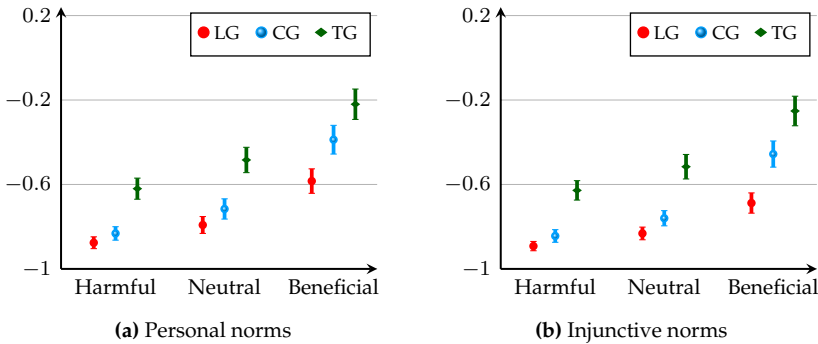


Figure 41: Personal and Injunctive norms for messages that do not benefit the sender.

Consistent with the results for the main messages reported in Subsection 4.3.4, the effects of lying and truth-telling on personal and injunctive norms for messages that do not benefit the sender mirror the ones on punishment. Only around 10% of observers consider a false report that benefits the receiver at the sender’s expense as somewhat inappropriate (or worse), either from a personal or a social perspective. This fraction increases to around 20% when the receiver is economically unaffected by this false message. When these messages are not false, they are considered somewhat appropriate (or better) by around 93% of observers. On

²10% of observers punish when the receiver is not worse off, whereas 20% do so if it affects the receiver.

³For exposition, the y-axis scale in Figure 41 is smaller than the full range in Figure 11.

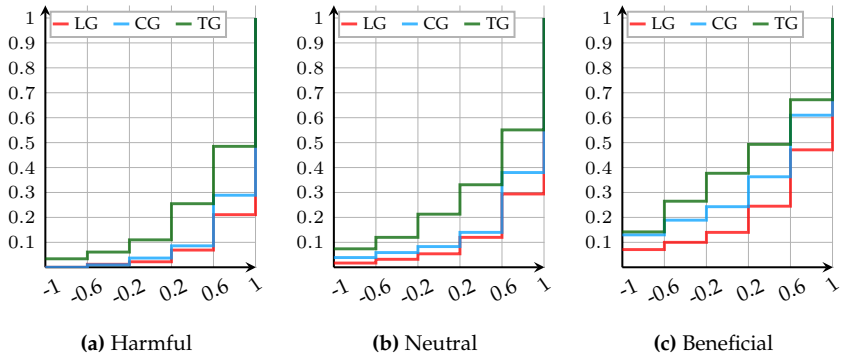


Figure 42: Empirical cumulative distribution functions of personal norms for messages that do not benefit the sender by game and situation.

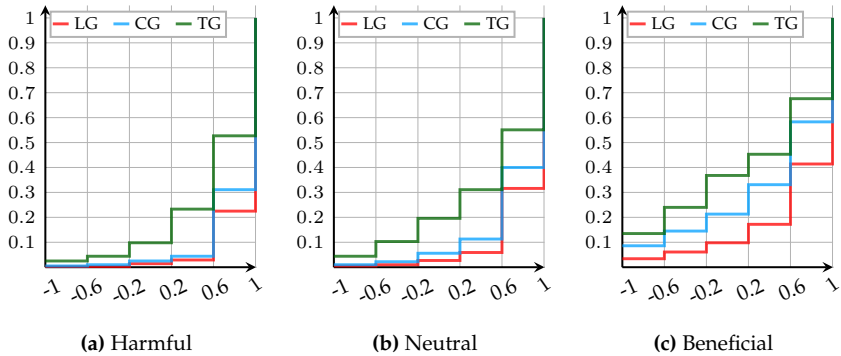


Figure 43: Empirical cumulative distribution functions of injunctive norms for messages that do not benefit the sender by game and situation.

the other hand, more than 60% of observers judge as appropriate (both personally and socially) a false report that harms the receiver but also the sender to the same extent.

Finally, we analyze whether there is a clear consensus about the message's appropriateness or multiple norms arise for these messages. Figures 44 and 45 show the histograms of personal and injunctive norms for the messages that do not benefit the sender.

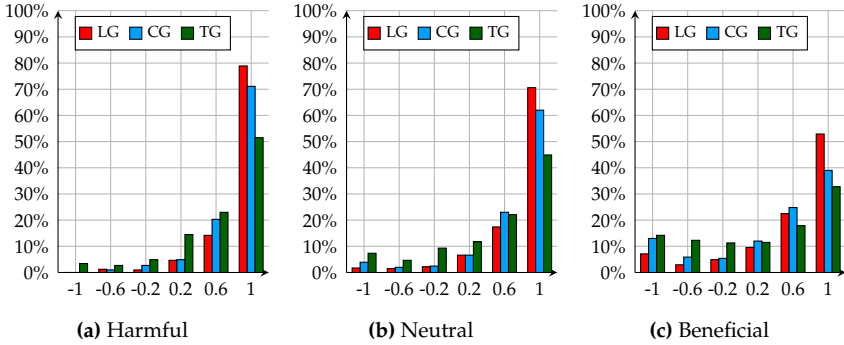


Figure 44: Distribution of personal norms of messages that do not benefit the sender.

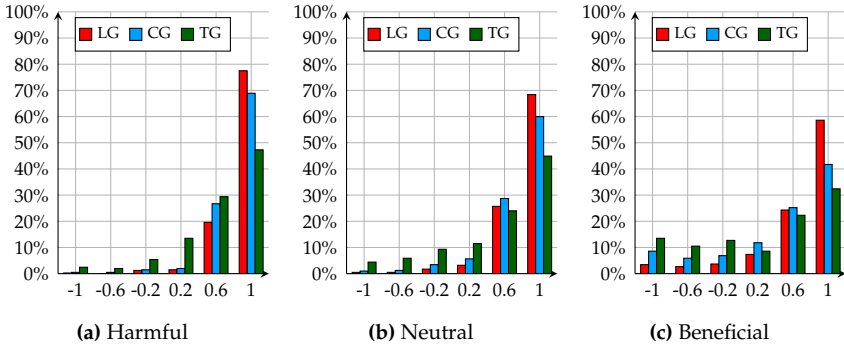


Figure 45: Distribution of injunctive norms of messages that do not benefit the sender.

In line with punishment results, we observe that messages that are not beneficial for the sender are considered mostly appropriate, either from a personal or a social perspective. It is noteworthy that a mutually harmful true message is considered somewhat appropriate (or better) by more than 85% of observers, that is, even if it makes the receiver worse-off, a message is appropriate if it decreases the sender's payoff to the same extent. Conversely, when this mutually harmful message is false, the strong consensus about its appropriateness disappears, although less than

25% of observers consider it inappropriate or worse. Still, more than 50% of observers consider it appropriate or better, reinforcing the role of distributional concerns on norms.

C.1.2 Norms ECDFs for messages that benefit the sender

Figures 46, 47, and 48 show the cumulative distribution functions of each norm for sender-benefiting messages across games and payoff situations.

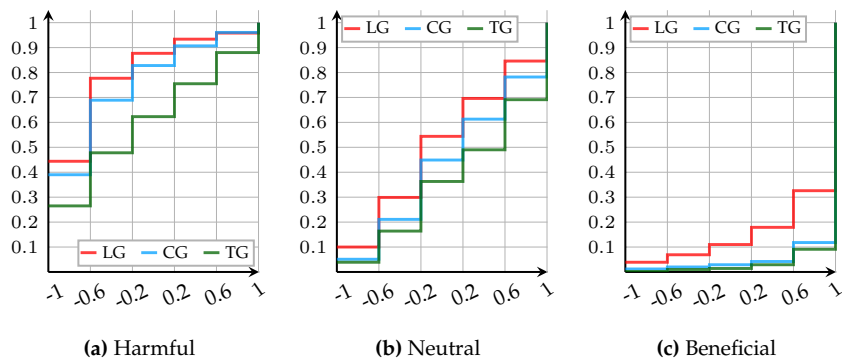


Figure 46: Empirical cumulative distribution functions (ECDFs) of personal norms for messages that benefit the sender by game and situation.

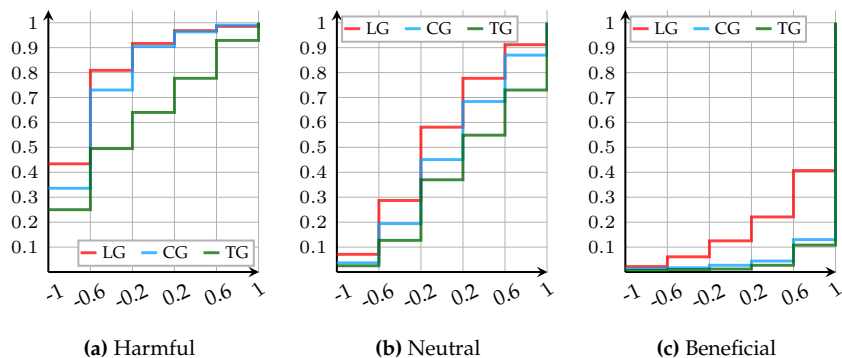


Figure 47: Empirical cumulative distribution functions (ECDFs) of injunctive norms for messages that benefit the sender by game and situation.

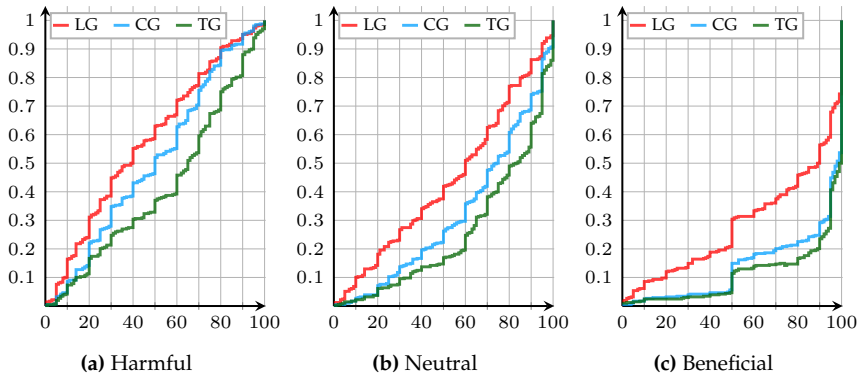


Figure 48: Empirical cumulative distribution functions (ECDFs) of descriptive norms for messages that benefit the sender by game and situation.

For each norm, the distributional comparison results are similar to those from centrality tests explained in Subsection 4.3.4. Across all payoff situations, the empirical distributions of personal and social norms for lies are dominated by the norms of their payoff-equivalent choices. Hence, all lies are judged as less appropriate and expected to be less frequent. On the other hand, the distributions for truthful reports dominate their payoff-equivalent choices, so truthful reports are judged as more appropriate and expected to be reported more frequently.

However, contrary to the results for punishment, harmful false reports differ in distribution from their payoff-equivalent choices ($p < 0.001$), whereas there is no statistically significant difference in the norm's distribution between beneficial truths and their payoff-equivalent choices ($p > 0.4$).

Finally, empirical expectations for harmful truths and beneficial lies align with the results of punishment, personal, and injunctive norms: half of observers expect that at least 90% of senders tell a beneficial lie. On the other hand, 75% of observers expect that at least 30% of senders tell a harmful truth.

C.2 Regression results on game-order effects

In this section, we examine whether the order of play affects punishment. Table 28 presents the regression results on overall order effects.

Dependent variable = Punishment	(1)	(2)
1 st CG	-0.113 (0.152)	-0.113 (0.152)
Reporting setting	0.127 (0.095)	
LG		0.279** (0.110)
TG		-0.026 (0.110)
1 st CG $\times$ Reporting setting	-0.095 (0.135)	
1 st CG $\times$ LG		-0.140 (0.155)
1 st CG $\times$ TG		-0.051 (0.155)
Harmful situation	1.591*** (0.078)	1.591*** (0.078)
Beneficial situation	-1.007*** (0.078)	-1.007*** (0.078)
S ₁₀	-2.793*** (0.063)	-2.793*** (0.063)
Intercept	3.450*** (0.121)	3.450*** (0.121)
Number of observations	7344	7344
Groups (number of participants)	408	408
<i>p</i> -value of Wald χ^2 test for order effects	0.280	0.411

Notes: Random effects regression results. Baseline categories: Setting order = Reporting first, setting or game = CG, situation = Neutral, and S₁₀ = 0. S₁₀ indicates the sender does *not* benefit from the message. Column (1) controls for the setting type: Reporting (LG and TG) or Choosing. Column (2) controls for the game. Standard errors in parentheses. 1st CG coefficients capture effects of Choosing first relative to Reporting first. LG and TG denote that the punishment is in the Lying and Truth games. ****p* < 0.01, ***p* < 0.05, **p* < 0.1.

Table 28: Regression results of the game-order analyses of punishment.

Consistent with the non-parametric results in Section 4.3, we find no significant order effects on punishment, either overall or when considering each game and setting separately.

C.2.1 Main regression estimations controlling for game-order effects

To examine whether our main lying and truth-telling effects are driven by order differences, we extend Model (4.1) in Subsection 4.3.3 to control for order effects. Specifically, we allow the order of settings (Reporting vs. Choosing first) to change behavior differently across games.

$$\begin{aligned}
 PUN_{ijgs} = & \alpha + \beta_1 S_{10_j} + \beta_2 LG_g + \beta_3 TG_g \\
 & + \beta_4 (S_{10_j} \times LG_g) + \beta_5 (S_{10_j} \times TG_g) \\
 & + \beta_6 (LG_g \times 1^{st} CG_i) + \beta_7 (TG_g \times 1^{st} CG_i) \quad (C.1) \\
 & + \beta_8 (S_{10_j} \times LG_g \times 1^{st} CG_i) \\
 & + \beta_9 (S_{10_j} \times TG_g \times 1^{st} CG_i) + \mu_i + \epsilon_{ijgs},
 \end{aligned}$$

As it can be noticed in Table 29, lying's aggravating effect on punishment is preserved on the aggregate after controlling for order effects. False reports are significantly more punished than their corresponding choices in both setting orders, and testing all restrictions jointly (all p -values < 0.001). However, discerning the effect for each situation separately, we find differences across order conditions depending on the lie type.

For harmful messages, the joint test indicates that lying increases punishment ($p = 0.003$), but its effect varies between order conditions: It is significant only when the CG is played first ($0.623, p < 0.001$), but there is no effect when the Reporting setting is played first ($-0.103, p = 0.610$). Moreover, as explained in Subsection 4.3.1 using non-parametric methods, the effect of lying on punishment for harmful reports does not hold for between-subject comparisons. Therefore, our results suggest that lying has no significant influence on punishment for harmful messages.

D.V: Punishment	Harmful	Neutral	Beneficial	All
LG	-0.103 (0.202)	0.784*** (0.196)	1.074*** (0.208)	0.585*** (0.145)
TG	-1.735*** (0.283)	-0.980*** (0.182)	-0.167* (0.100)	-0.961*** (0.142)
S ₁₀	-6.706*** (0.279)	-2.701*** (0.203)	0.368* (0.171)	-3.013*** (0.155)
LG × S ₁₀	0.186 (0.243)	-0.863*** (0.244)	-1.157*** (0.262)	-0.611*** (0.181)
TG × S ₁₀	2.495*** (0.382)	1.721*** (0.286)	1.392*** (0.247)	1.869*** (0.246)
LG × 1 st CG	0.725** (0.273)	-0.093 (0.265)	-1.098*** (0.259)	-0.155 (0.180)
TG × 1 st CG	0.877* (0.387)	0.824** (0.266)	-0.147 (0.190)	0.518** (0.196)
LG × S ₁₀ × 1 st CG	-0.853** (0.322)	0.078 (0.316)	0.868** (0.326)	0.031 (0.221)
TG × S ₁₀ × 1 st CG	-1.245* (0.496)	-1.225** (0.375)	-0.941** (0.319)	-1.137*** (0.303)
Mean punishment baseline	7.180	3.340	0.750	3.757
Number of observations	2448	2448	2448	7344
Adj. R ²	0.622	0.345	0.058	0.212

Notes: Fixed-effects regression results with clustered standard errors at the participant level (in parentheses). Baseline category is CG if S₁₀ = 0 and setting order = Reporting first. 1st CG coefficients capture effects of Choosing first relative to Reporting first. LG and TG denote that the punishment is in the Lying and Truth games. S₁₀ indicates that the sender does *not* benefit from the message sent. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 29: Fixed-effects regression results controlling for order effects.

By contrast, although the aggravating effect of beneficial lying is jointly significant ($p < 0.001$), it appears just when the lie is punished first (1.074, $p < 0.001$). Conversely, when the false report is punished after the payoff equivalent choice, there are no significant differences (-0.025, $p = 0.873$). However, our non-parametric comparisons show that the effect of lying in beneficial messages is preserved for the between-subject comparison.

Finally, there is no order difference for neutral lies: lying is always a significant aggravating factor on punishment (0.784 for Reporting first and 0.691 for Choosing first, $p < 0.001$ for both order conditions).

On the other hand, the mitigating effect of truth-telling is more robust. Pooling all situations, there is a significant decrease in punishment both when the truth is punished first (-0.961 , $p < 0.001$), and second (-0.443 , $p = 0.001$). When discerning for each situation, the joint tests indicate significant mitigating effects in all situations, although the strength of the effect differs across types.

For harmful reports, truth-telling reduces punishment under both setting orders: if the truth is punished first, the effect is -1.735 , whereas it is -0.858 if it is sanctioned second ($p < 0.001$ for both). For neutral truths, the mitigating effect appears when the truth is sanctioned first (-0.980 , $p < 0.001$), but it is not statistically significant when the choice is sanctioned before (-0.157 , $p = 0.418$). For beneficial truths, both are marginally significant and the differences between order conditions are smaller: an effect of -0.167 when the truth is sanctioned first ($p = 0.096$), and -0.314 when it is sanctioned second ($p = 0.052$). However, as shown in Subsection 4.3.2, our non-parametric comparisons show that the effect of truth-telling in harmful and neutral messages is preserved for the between-subject comparison.

Therefore, our main punishment findings are robust to the order effects given by the within-subject design: While lying's aggravating effect is relevant only in the absence of consequential violations, truth-telling significantly lowers punishment for selfish reports, although it does not eliminate it entirely. Hence, there is an asymmetry in the mitigating effect of truth-telling and the aggravating role of dishonesty in punishment.

C.3 Punishment model with situations interacted

To estimate the differences in punishment driven by lying and truth-telling across payoff situations, we use the following model that expands

Model (4.1) by adding situation interactions.

$$\begin{aligned}
PUN_{ijgs} = & \alpha + \beta_1 S_{10j} + \beta_2 LG_g + \beta_3 TG_g + \beta_4 H_s + \beta_5 B_s \\
& + \beta_6 (S_{10j} \times LG_g) + \beta_7 (S_{10j} \times TG_g) \\
& + \beta_8 (S_{10j} \times H_s) + \beta_9 (S_{10j} \times B_s) \\
& + \beta_{10} (LG_g \times H_s) + \beta_{11} (LG_g \times B_s) \\
& + \beta_{12} (TG_g \times H_s) + \beta_{13} (TG_g \times B_s) \\
& + \beta_{15} (S_{10j} \times LG_g \times H_s) + \beta_{15} (S_{10j} \times LG_g \times B_s) \\
& + \beta_{16} (S_{10j} \times TG_g \times H_s) + \beta_{17} (S_{10j} \times TG_g \times B_s) \\
& + \mu_i + \epsilon_{ijgs},
\end{aligned} \tag{C.2}$$

where H_s is a dummy variable equal to 1 if being in the harmful situation, B_s is the analog for the beneficial situation, and the baseline is the neutral situation. The rest of variables are those in Model (4.1).

Table 30 contains the estimated differences between situations of the aggravating effect of lying and the mitigating influence of truth-telling. Table 31 provides the results of the estimated regression of Model (C.2), controlling for observer heterogeneity with fixed effects.

Differences in the effects on punishment	Lying (aggravating)	Truth-telling (mitigating)
Harmful vs. Neutral	-0.478***	-0.728***
Neutral vs. Beneficial	0.213	-0.328**
Harmful vs. Beneficial	-0.265	-1.056***

Table 30: Differences between payoff situations of the effects of lying and truth-telling on punishment for sender-benefiting messages. Estimates come from a fixed-effects regression with standard errors clustered at the participant level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

D.V: Punishment

LG	0.738*** (0.132)
TG	-0.569*** (0.134)
Harmful	3.694*** (0.164)
Beneficial	-2.532*** (0.136)
S ₁₀	-2.745*** (0.149)
LG × Harmful	-0.478*** (0.177)
TG × Harmful	-0.728*** (0.189)
LG × Beneficial	-0.213 (0.148)
TG × Beneficial	0.328** (0.160)
LG × S ₁₀	-0.824*** (0.148)
TG × S ₁₀	1.108*** (0.190)
Harmful × S ₁₀	-3.850*** (0.186)
Beneficial × S ₁₀	3.002*** (0.169)
LG × Harmful × S ₁₀	0.583*** (0.206)
TG × Harmful × S ₁₀	0.765*** (0.222)
LG × Beneficial × S ₁₀	0.100 (0.186)
TG × Beneficial × S ₁₀	-0.186 (0.207)
Number of observations	7344
Adj. R ²	0.494

Notes: Fixed-effects regressions with standard errors clustered at the participant level (in parentheses). Baseline categories are CG, neutral situation, and S₁₀ = 0.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 31: Results of the regression estimation of Model (C.2).

C.4 Regressions results of norms analyses

In this section, we provide the regression results on the effect of lying and truth-telling on each norm measure explained in Subsection 4.3.4. For personal and injunctive social norms, we use Model (C.3), which mirrors Model (4.1) but uses the assessment of the message as the dependent variable.

$$\begin{aligned} Norm_{ijgs} = & \alpha + \beta_1 S_{10_j} + \beta_2 LG_g + \beta_3 TG_g \\ & + \beta_4 (S_{10_j} \times LG_g) + \beta_5 (S_{10_j} \times TG_g) + \mu_i + \epsilon_{ijgs} \end{aligned} \quad (C.3)$$

where $Norm_{ijgs}$ denotes observer i 's assessment (either the personal norm or the injunctive social norm) in game g and situation s if the sender reports message j . The rest of variables are those in Model (4.1).

D.V: Personal norms	Harmful	Neutral	Beneficial	All
LG	-0.086*** (0.021)	-0.142*** (0.026)	-0.201*** (0.026)	-0.146*** (0.018)
TG	0.310*** (0.031)	0.143*** (0.023)	0.029* (0.014)	0.161*** (0.016)
S_{10}	1.341*** (0.037)	0.558*** (0.046)	-0.525*** (0.036)	0.458*** (0.031)
$LG \times S_{10}$	0.129*** (0.029)	0.229*** (0.035)	0.400*** (0.042)	0.253*** (0.027)
$TG \times S_{10}$	-0.520*** (0.049)	-0.375*** (0.040)	-0.195*** (0.031)	-0.363*** (0.032)
Mean baseline PN	-0.510	0.148	0.912	0.187
Number of observations	2448	2448	2448	7344
Adj. R^2	0.589	0.083	0.112	0.077

Notes: Fixed-effects regressions with standard errors clustered at the participant level (in parentheses). Norm values are scaled from -1 (very personally inappropriate) to 1 (very personally appropriate). Baseline category is CG when $S_{15} = 1$. S_{10} indicates that the sender does not benefit from the message. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 32: Personal norms regressions by situation and for the pooled sample.

D.V: Injunctive norms	Harmful	Neutral	Beneficial	All
LG	-0.075*** (0.019)	-0.147*** (0.027)	-0.241*** (0.029)	-0.148*** (0.018)
TG	0.333*** (0.030)	0.174*** (0.025)	0.027* (0.014)	0.178*** (0.017)
S ₁₀	1.412*** (0.030)	0.654*** (0.038)	-0.450*** (0.033)	0.539*** (0.025)
LG × S ₁₀	0.124*** (0.027)	0.230*** (0.035)	0.473*** (0.043)	0.275*** (0.026)
TG × S ₁₀	-0.546*** (0.048)	-0.416*** (0.044)	-0.233*** (0.030)	-0.398*** (0.034)
Mean baseline IN	-0.570	0.106	0.907	0.148
Number of observations	2448	2448	2448	7344
Adj. R ²	0.675	0.214	0.101	0.134

Notes: Fixed-effects regressions with standard errors clustered at the participant level (in parentheses). Baseline category is CG when S₁₅ = 1. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 33: Injunctive norms regressions by situation and for the pooled sample.

Since the two expected frequencies in a situation sum to 100, we model only the descriptive norms for senders' payoff-maximizing messages, denoted DN_{igs} , using the reduced specification below.

$$DN_{igs} = \alpha + \beta_1 LG_g + \beta_2 TG_g + \mu_i + \epsilon_{igs}, \quad (C.4)$$

D.V: Descriptive norms	Harmful	Neutral	Beneficial	All
LG	-6.885*** (1.206)	-12.000*** (1.495)	-14.485*** (1.606)	-11.123*** (1.081)
TG	9.679*** (1.383)	6.436*** (1.248)	2.532* (1.026)	6.216*** (0.838)
Mean baseline DN	49.172	68.502	86.502	68.059
Number of observations	1224	1224	1224	3672
Adj. R ²	0.106	0.128	0.121	0.066

Notes: Fixed-effects regressions with standard errors clustered at the participant level (in parentheses). Baseline = CG. LG and TG capture lying and truth-telling effects.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 34: Descriptive norms regressions by situation and for the pooled sample.

C.5 Reduced models of punishment and norms

Below, we report results from reduced versions of Model (4.2) including one or two norms.

D.V: Punishment	PN only	IN only	DN only	PN + IN	PN + DN	IN + DN
PN	1.162*** (0.035)			0.633*** (0.029)	1.214*** (0.018)	
IN		1.199*** (0.035)		0.670*** (0.029)		1.229*** (0.018)
DN			-0.064*** (0.003)		0.010*** (0.001)	0.007*** (0.001)
N	7344	7344	3672	7344	7344	7344
Adj. R ²	0.359	0.361	0.143	0.404	0.372	0.368
AIC	34114.5	34085.7	18373.5	33584.5	33967.0	34012.8
BIC	34128.4	34099.5	18385.9	33605.2	33987.7	34033.5

Notes: Fixed-effects regression results for the pooled sample with clustered standard errors at the participant level (in parentheses). PN and IN are scaled from 0 (very appropriate) to 5 (very inappropriate). DN is scaled from 0 to 100. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 35: Fixed-effects regressions of punishment on norms: pooled sample.

D.V: Punishment	PN only	IN only	DN only	PN + IN	PN + DN	IN + DN
PN	1.549*** (0.040)			0.953*** (0.056)	1.561*** (0.030)	
IN		1.520*** (0.043)		0.690*** (0.056)		1.520*** (0.030)
DN			-0.032*** (0.005)		0.004*** (0.002)	0.000 (0.002)
N	2448	2448	1224	2448	2448	2448
Adj. R ²	0.499	0.467	-0.399	0.533	0.500	0.467
AIC	11469.8	11619.8	5282.4	11298.7	11465.7	11621.8
BIC	11481.4	11631.4	5292.7	11316.2	11483.1	11639.2

Notes: Fixed-effects regression with clustered standard errors at the participant level (in parentheses). PN and IN are scaled from 0 (very appropriate) to 5 (very inappropriate). DN is scaled from 0 to 100. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 36: Fixed-effects regressions of punishment on norms: harmful situation.

D.V: Punishment	PN only	IN only	DN only	PN + IN	PN + DN	IN + DN
PN	0.908*** (0.044)			0.511*** (0.044)	0.934*** (0.030)	
IN		0.964*** (0.043)		0.561*** (0.046)		0.971*** (0.031)
DN			-0.019*** (0.004)		0.012*** (0.001)	0.009*** (0.001)
N	2448	2448	1224	2448	2448	2448
Adj. R ²	0.171	0.175	-0.440	0.227	0.202	0.194
AIC	10293.5	10281.2	4753.5	10124.4	10200.4	10226.6
BIC	10305.1	10292.8	4763.8	10141.8	10217.8	10244.0

Notes: Fixed-effects regression with clustered standard errors at the participant level (in parentheses). PN and IN are scaled from 0 (very appropriate) to 5 (very inappropriate). DN is scaled from 0 to 100. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 37: Fixed-effects regressions of punishment on norms: neutral situation.

D.V: Punishment	PN only	IN only	DN only	PN + IN	PN + DN	IN + DN
PN	0.397*** (0.056)			0.260*** (0.044)	0.389*** (0.036)	
IN		0.405*** (0.055)		0.202*** (0.047)		0.377*** (0.038)
DN			-0.012*** (0.004)		0.000 (0.001)	-0.002 (0.001)
N	2448	2448	1224	2448	2448	2448
Adj. R ²	-0.107	-0.116	-0.467	-0.098	-0.108	-0.114
AIC	9712.2	9731.6	4354.4	9692.3	9714.0	9731.0
BIC	9723.8	9743.2	4364.6	9709.7	9731.4	9748.4

Notes: Fixed-effects regression results for the beneficial situation with clustered standard errors at the participant level (in parentheses). PN and IN are scaled from 0 (very appropriate) to 5 (very inappropriate). DN is scaled from 0 to 100. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 38: Fixed-effects regressions of punishment on norms: beneficial situation.

C.6 Comparisons norms of observers and judges

Table 39 provides the norms comparison between the judges and the entire observer samples.

Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	-0.60 (0.52)	-0.36 (0.58)	-0.24***
CG	-0.51 (0.52)	-0.28 (0.62)	-0.23***
TG	-0.20 (0.72)	0.16 (0.66)	-0.36***
(a) Personal norms – Harmful			
Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	-0.65 (0.46)	-0.47 (0.55)	-0.17***
CG	-0.57 (0.41)	-0.47 (0.51)	-0.10*
TG	-0.24 (0.67)	-0.00 (0.69)	-0.23***
(d) Injunctive norms – Harmful			
Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	42.29 (27.88)	44.59 (26.31)	-2.30
CG	49.17 (26.64)	50.41 (26.02)	-1.24
TG	58.85 (28.55)	60.03 (27.11)	-1.18
(g) Descriptive norms – Harmful			

Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	0.01 (0.67)	0.19 (0.62)	-0.19***
CG	0.16 (0.62)	0.45 (0.58)	-0.29***
TG	0.30 (0.62)	0.62 (0.47)	-0.31***
(b) Personal norms – Neutral			
Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	-0.05 (0.52)	0.07 (0.55)	-0.12**
CG	0.11 (0.57)	0.26 (0.55)	-0.16**
TG	0.28 (0.62)	0.47 (0.51)	-0.19***
(e) Injunctive norms – Neutral			
Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	56.50 (29.99)	57.73 (30.46)	-1.23
CG	68.50 (26.54)	71.46 (27.73)	-2.95*
TG	74.94 (25.25)	77.15 (25.47)	-2.21
(h) Descriptive norms – Neutral			

Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	0.71 (0.52)	0.56 (0.58)	0.15***
CG	0.91 (0.31)	0.93 (0.26)	-0.02
TG	0.94 (0.21)	0.89 (0.29)	0.05***
(c) Personal norms – Beneficial			
Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	0.67 (0.52)	0.49 (0.58)	0.18***
CG	0.91 (0.31)	0.93 (0.29)	-0.02
TG	0.93 (0.26)	0.89 (0.29)	0.05***
(f) Injunctive norms – Beneficial			
Game	Observers <i>N</i> = 408	Judges <i>N</i> = 204	Mean O - J
LG	72.02 (31.69)	72.71 (28.60)	-0.69
CG	86.50 (23.03)	88.56 (22.63)	-2.06
TG	89.03 (21.28)	88.08 (21.90)	0.95
(i) Descriptive norms – Beneficial			

Table 39: Comparison of norms with and without previously elicited punishment, using all observer decisions, as graphically represented in Figure 14. Tables (a)–(c) report personal norms, Tables (d)–(f) injunctive norms, and Tables (g)–(i) descriptive norms. The first columns report means and standard deviations (in parentheses). The last columns report mean differences computed as observers minus judges, with significance based on two-sided Mann–Whitney tests. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

C.6.1 Comparisons using only the first setting elicited

We present the analogous between-subject comparisons from Subsection 4.3.4 using only responses from the first setting played. Hence, the observers and judges samples have 204 observations, ensuring that their norm elicitation occurs without prior elicitation in other games.

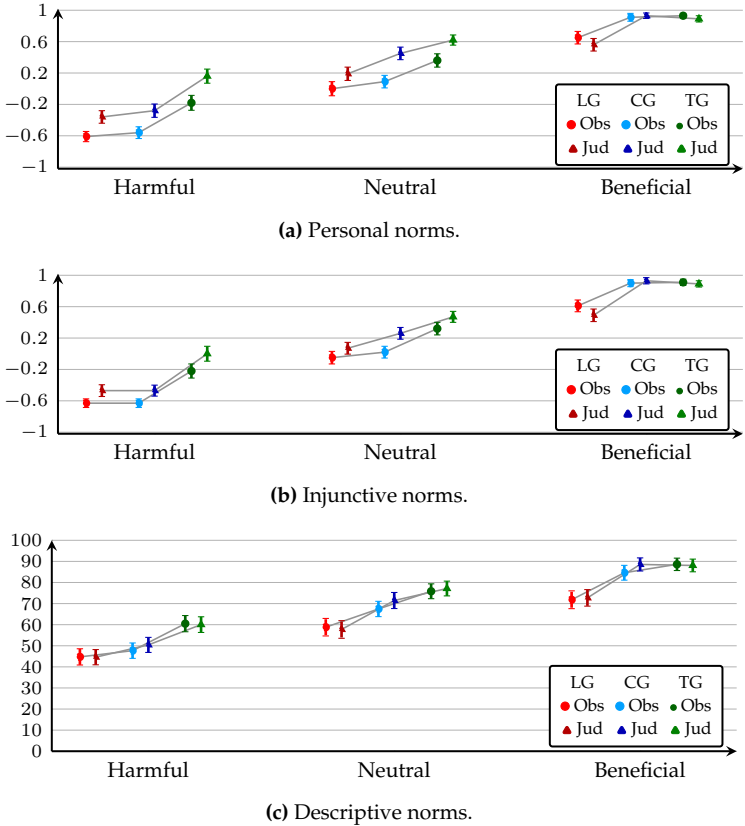


Figure 49: Comparison of norms with and without previous punishment elicitation using observations of the first setting elicited accordingly. Means and 95% confidence intervals are shown for observers (with punishment) and judges (without punishment) across situations and games. The top, middle, and bottom panels display personal (PN), injunctive (IN), and descriptive norms (DN).

Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	-0.61 (0.47)	-0.36 (0.58)	-0.25***
CG	-0.56 (0.55)	-0.28 (0.62)	-0.28***
TG	-0.18 (0.69)	0.16 (0.66)	-0.34***
(a) Personal norms – Harmful			
Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	-0.63 (0.40)	-0.47 (0.55)	-0.16**
CG	-0.63 (0.40)	-0.47 (0.51)	-0.16***
TG	-0.22 (0.66)	-0.00 (0.69)	-0.22***
(d) Injunctive norms – Harmful			
Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	44.73 (28.02)	44.59 (26.31)	0.14
CG	47.68 (26.56)	50.41 (26.02)	-2.73
TG	60.49 (27.98)	60.03 (27.11)	0.46
(g) Descriptive norms – Harmful			
Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	-0.00 (0.66)	0.19 (0.62)	-0.20**
CG	0.09 (0.58)	0.45 (0.58)	-0.36***
TG	0.36 (0.62)	0.62 (0.47)	-0.25***
(b) Personal norms – Neutral			
Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	-0.05 (0.58)	0.07 (0.55)	-0.12*
CG	0.02 (0.55)	0.26 (0.55)	-0.25***
TG	0.32 (0.58)	0.47 (0.51)	-0.15*
(e) Injunctive norms – Neutral			
Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	71.82 (30.75)	72.71 (28.60)	-0.88
CG	84.62 (25.54)	88.56 (22.63)	-3.95
TG	88.61 (21.39)	88.08 (21.90)	0.52
(h) Descriptive norms – Neutral			
Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	0.65 (0.58)	0.56 (0.58)	0.09**
CG	0.91 (0.36)	0.93 (0.26)	-0.02
TG	0.93 (0.22)	0.89 (0.29)	0.04**
(c) Personal norms – Beneficial			
Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	0.61 (0.55)	0.49 (0.58)	0.12***
CG	0.90 (0.33)	0.93 (0.29)	-0.02
TG	0.91 (0.29)	0.89 (0.29)	0.02
(f) Injunctive norms – Beneficial			
Game	Observers <i>N</i> = 204	Judges <i>N</i> = 204	Mean O - J
LG	58.81 (30.21)	57.73 (30.46)	1.08
CG	67.45 (26.45)	71.46 (27.73)	-4.00*
TG	75.84 (25.65)	77.15 (25.47)	-1.31
(i) Descriptive norms – Beneficial			

Table 40: Comparison of norms with and without previously elicited punishment, using observations of the first elicited setting, as graphically represented in Figure 49. Tables (a)–(c) report personal norms, Tables (d)–(f) injunctive norms, and Tables (g)–(i) descriptive norms. The first columns report means and standard deviations (in parentheses). The last columns report mean differences, computed as observers minus judges, with significance based on two-sided Mann–Whitney tests. Reported standard deviations are approximated from the 95% confidence intervals. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

C.7 Experimental procedure diagrams

The diagrams in Figures 50, and 51 represent the procedure followed by senders and receivers in both order conditions. The procedures are explained in further detail in Subsection 4.2.5.

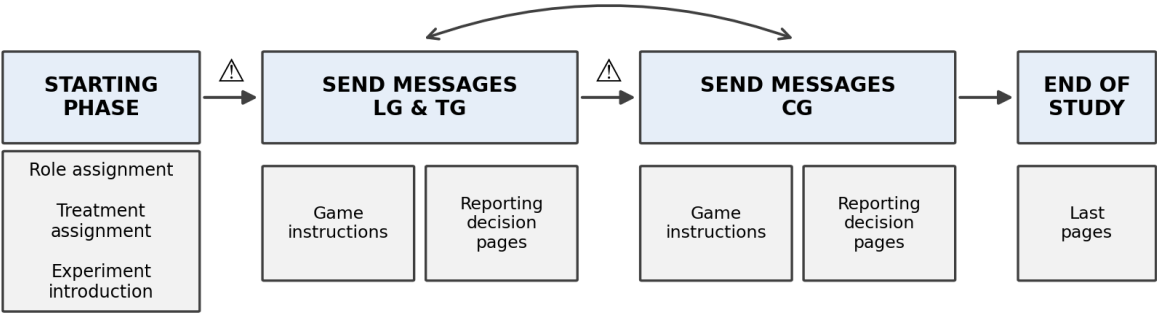
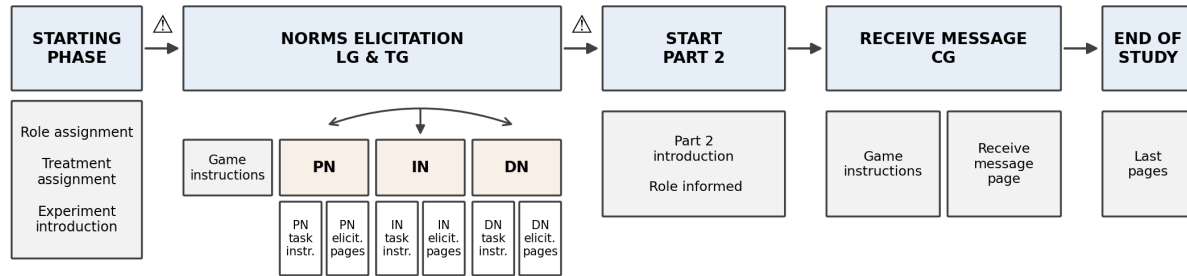
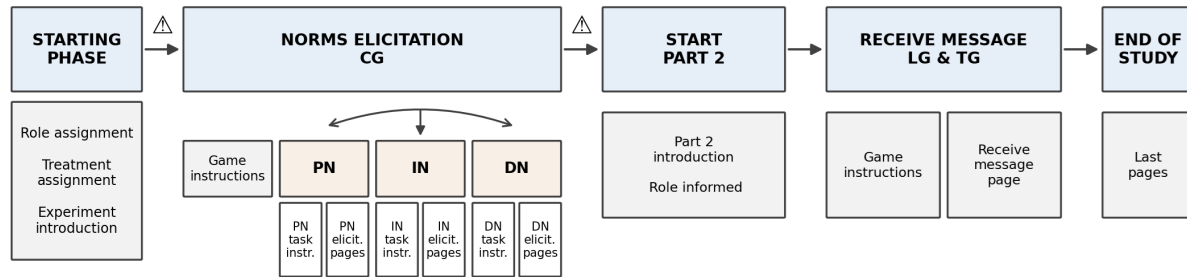


Figure 50: Procedure of senders in the Reporting first game-order condition. Warning icons indicate where attention checks were conducted. Curved arrows represent the randomization of settings' order.



(a) Procedure of receivers in the Choosing Game. They were elicited norms in the Lying and Truth games.



(b) Procedure of receivers in the Lying and Truth games. They were elicited norms in the Choosing Game.

Figure 51: Procedures for receivers in the between-subject order conditions. Warning icons indicate where attention checks were conducted. Curved arrows represent the randomization of the order within the norms-elicitation tasks.

C.8 Sociodemographic characteristics

Table 41 reports descriptive statistics on the sociodemographic characteristics of participants by role. As commented in Section 4.2.5, for each role, the sample was evenly split by gender (50% men and 50% women). The average age of the sample was 32.63 years (s.d. 8.03) for observers, 33.83 years (s.d. 7.91) for senders, and 32.06 years (s.d. 8.09) for receivers.

Variable	Category	Observers		Senders		Receivers	
		N	%	N	%	N	%
Religion	Non Religious	213	52.2%	227	55.6%	233	57.1%
	Christianity	128	31.4%	121	29.7%	127	31.1%
	Other	67	16.4%	60	14.7%	48	11.8%
Race/Ethnicity	Asian	37	9.1%	29	7.1%	34	8.3%
	Black	30	7.4%	21	5.1%	36	8.8%
	White	310	76.0%	326	79.9%	295	72.3%
	Other	31	7.6%	32	7.8%	43	10.5%
Student status	No	318	77.9%	343	84.1%	312	76.5%
	Yes	90	22.1%	65	14.9%	96	23.5%
Employment status	Full-Time	210	51.5%	234	57.4%	180	44.1%
	Unemployed	58	14.2%	38	9.3%	74	18.1%
	Part-Time	80	19.6%	62	14.2%	78	19.1%
	Not in paid work	24	5.9%	38	9.3%	41	10.0%
	Other	36	8.8%	36	8.8%	35	8.6%
First language	English	363	89.0%	379	92.9%	370	90.7%
	Other	45	11.0%	29	7.1%	38	9.3%
Country of residence	United Kingdom	273	66.9%	246	60.3%	191	46.8%
	United States	135	33.1%	162	39.7%	217	53.2%
Country of birth	United Kingdom	227	55.6%	207	50.7%	147	36.0%
	United States	121	29.7%	147	38.5%	211	51.7%
	Other	60	14.7%	44	10.8%	50	12.3%
Boycotting	No	145	35.5%	165	40.4%	142	37.3%
	Yes	263	64.5%	243	59.6%	256	62.7%

Table 41: Sociodemographic characteristics of the observer, sender, and receiver participant pools.

C.9 Instructions

Since the experiment involved three roles, the full set of instructions is extensive. To keep this section manageable, we include only the instructions and decision screens for each setting. These are the most relevant pages for understanding the games and the outcomes analyzed in the paper. The instructions for the norm-elicitation tasks are similar for all participants, so, to avoid repetition, we display only those for observers. For receivers, we include only the explanation of each part, the instructions for both settings, and examples of the pages displaying the received message. The full set of instructions for all roles is available in the OSF repository.

C.9.1 Observers' instructions

We first present observers' instructions for each game in the first part, where punishment was elicited. We then show the norm-elicitation tasks for the Reporting setting only, as those for the Choosing game are analogous. The figures below follow the order of the Reporting-first condition.

Part 1: Instructions for your task in this part.

In this part of the study, each participant is assigned the role of either Person 1, Person 2, or Person 3. **You will keep the same role in all scenarios in this part.** Your role determines which decisions you can make and what payments you can get. The decisions you make in one part of the study will not affect your decisions or possible earnings in any other part. Below, you can see your role.

Your role:

Person 3

In this part of the study, you will **observe some scenarios** that involve other randomly selected participants. **For each scenario, your task as Person 3 is to make decisions in several situations** that might affect other participants' earnings. Your decisions cannot be reversed, so it is important that you **read the instructions for each scenario carefully**.

If this part is used to determine participants' payments, a computer will **randomly select one single situation and will randomly match you with other participants**. Decisions in one situation will not influence the rest of the situations. Hence, **you should consider each situation as if it were the only one selected for payments**.

Your **payment** in this part is the **flat rate** you are paid for all your decisions. This payment will **not be affected by your decisions or those of others** in any part of the study. **Your earnings in this study are greater** than those of participants in other roles. Keep in mind that **you must complete the entire study to receive any payment**.

After reading the prior guidelines, press the 'Next' button to see the first scenario and your task in Part 1 of the study.

Next

Figure 52: Instructions of the first part (punishment elicitation) for observers.

Part 1: Instructions of Scenario 1.

Now, you will observe some **situations of Scenario 1**, which involves participants in the role of Person 1, 2 and 3. To determine your payments if Part 1 is used, a computer will randomly select one single situation and randomly match you with two other participants. As **Person 3, your role in this part is to decide whether to deduct points from Person 1** in each situation.

Scenario 1

For each situation, there are **two possible outcomes that Person 1 could observe**. A priori, both outcomes are equally likely. These outcomes are a filled square (left) and a striped square (right).



At the beginning of each situation, **Person 1 will be shown the outcome in that situation**. For example, if the outcome of the situation is a filled square, Person 1 will observe the following:

The outcome in this situation is 

In this scenario, **Person 1 is asked to report the observed outcome of the situation to Person 2**, who does not know it. **Person 1's report** will determine the **payoffs** for both Person 1 and Person 2 in each situation. Before deciding what outcome to report, **Person 1 observes all payoffs** associated with each possible report in a table similar to the one presented below.

	Person 1's payoff	Person 2's payoff
If Person 1 reports 	##	##
If Person 1 reports 	##	##

Along with Person 1's message, **Person 2 will only be informed about their own payoff according to Person 1's report**. Hence, Person 2 will not know the observed payoff table corresponding to the message received nor the outcome Person 1 had to report.

Your task:

You have to **decide whether to deduct points from Person 1's payoff for each possible report they could make** in each situation. To do so, you have to select a number (from 0 to 10) for each of the two possible reports that Person 1 can make. There is no right or wrong answer in this part of the study. Also, note that your answers in this part will not affect your earnings and will only be used to deduct points in the selected situation, not being shown to any other participant in this study.

Participants in the role of Person 1 are aware of your role in the scenario. They know you have to decide whether to deduct up to 10 points from their payment for each potential report they could make. They are also aware that you will never see their actual report in any scenario. For more details on Person 1's instructions, you can see them in a separate tab by clicking [here](#).

After carefully reading all the previous instructions, press the 'Next' button to start your task in Scenario 1.



Figure 53: Instructions of the Reporting setting for observers in the punishment part.

The **outcome that Person 1 observed** in this situation is:



The table below shows both participants' payoffs corresponding to each **possible report** Person 1 could make.

	Person 1's payoff	Person 2's payoff
If Person 1 reports 	10	10
If Person 1 reports 	15	5

Please, select below how many **points (0-10)** you would like to deduct from Person 1's earnings for each **possible report**. These points will be deducted if Person 1 makes this report.

	0	1	2	3	4	5	6	7	8	9	10
If Person 1 reports 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
If Person 1 reports 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Please, press the 'Next' button after completing this page (entering a value for each possible report).

Next

To display the instructions of this task again, click [here](#).

Figure 54: Decision page in the Lying Game (harmful situation) from the observer's perspective.

The **outcome that Person 1 observed** in this situation is:



The table below shows both participants' payoffs corresponding to each **possible report** Person 1 could make.

	Person 1's payoff	Person 2's payoff
If Person 1 reports 	10	10
If Person 1 reports 	15	5

Please, select below how many **points (0-10)** you would like to deduct from Person 1's earnings for each **possible report**. These points will be deducted if Person 1 makes this report.

	0	1	2	3	4	5	6	7	8	9	10
If Person 1 reports 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
If Person 1 reports 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Please, press the 'Next' button after completing this page (entering a value for each possible report).

Next

To display the instructions of this task again, click [here](#).

Figure 55: Decision page in the Truth Game (harmful situation) from the observer's perspective.

Part 1: Instructions of Scenario 2.

Now, you will observe some **situations of a new scenario, Scenario 2**, which involves participants in the role of Person 1, 2 and 3. To determine your payments if Part 1 is used, a computer will randomly select one single situation and randomly match you with two other participants. As **Person 3, your role in this part is to decide whether to deduct points from Person 1** in each situation.

Scenario 2

For each situation, there are **two possible outcomes that Person 1 could choose**. These outcomes are a filled circle (left) and a striped circle (right).



In this scenario, **Person 1 is asked to choose the outcome of the situation** and report it to Person 2.

Person 1's choice will determine the **payoffs** for both Person 1 and Person 2 in each situation. Before deciding what outcome to choose, **Person 1 observes all payoffs** associated with each possible choice in a table similar to the one presented below.

	Person 1's payoff	Person 2's payoff
If Person 1 chooses 	##	##
If Person 1 chooses 	##	##

Along with Person 1's message, **Person 2 will only be informed about their own payoff according to Person 1's choice**. Hence, Person 2 will not know the observed payoff table corresponding to the message received.

Your task:

You have to **decide whether to deduct points from Person 1's payoff for each possible choice they could make** in each situation. To do so, you have to select a number (from 0 to 10) for each of the two possible choices that Person 1 can make. There is no right or wrong answer in this part of the study. Also, note that your answers in this part will not affect your earnings and will only be used to deduct points in the selected situation, not being shown to any other participant in this study.

Participants in the role of Person 1 are aware of your role in the scenario. They know you have to decide whether to deduct up to 10 points from their payment for each potential choice they could make. They are also aware that you will never see their actual choice in any scenario. For more details on Person 1's instructions, you can see them in a separate tab by clicking [here](#).

After carefully reading all the previous instructions, press the 'Next' button to start your task in Scenario 2.

Next

Figure 56: Instructions of the Choosing Game for observers in the Reporting-first condition in the punishment part.

The table below shows both participants' payoffs corresponding to each possible choice Person 1 could make.

	Person 1's payoff	Person 2's payoff
If Person 1 chooses 	10	10
If Person 1 chooses 	15	5

Please, select below how many points (0-10) you would like to deduct from Person 1's earnings for each possible choice. These points will be deducted if Person 1 makes this choice.

	0	1	2	3	4	5	6	7	8	9	10
If Person 1 chooses 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
If Person 1 chooses 	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Please, press the 'Next' button after completing this page (entering a value for each possible choice).

Next

To display the instructions of this task again, click [here](#).

Figure 57: Decision page in the Choosing Game (harmful situation) from the observer's perspective.

Part 2: Instructions for your task in this part.

You have completed Part 1 of this study. Now, you start with Part 2.

In this part, you will observe some scenarios that involve other randomly selected participants. For each scenario, your task is to evaluate decisions made in several situations that might affect other participants' earnings. The evaluations you make in this part of the study will not affect your decisions or possible earnings in any other part and cannot be reversed.

If this part is used to determine participants' payments, a computer will randomly select one single situation and will randomly match you with other participants. Therefore, it is important that you read the instructions for each scenario and the evaluations you will make carefully. Your evaluations in one situation will not influence the rest of the situations. Hence, you should consider each situation as if it were the only one selected for payments.

Your payment in this part will depend on your evaluations. This payment will not be affected by your decisions or those of others in any other part of the study. Your earnings in this study are greater than those of participants in other roles. Keep in mind that you must complete the entire study to receive any payment.

After reading the prior guidelines, press the 'Next' button to see the first scenario and your task in Part 2 of the study.

Next

Figure 58: Instructions norms part for observers.

Part 2: Instructions of Scenario 1.

Now, you will observe some **situations of Scenario 1**, which involves participants in the role of Person 1, Person 2 and Person 3. To determine your payments if Part 2 is used, a computer will randomly select one single situation. In this part of the study, you have to **evaluate the decisions that Person 1 could make** in each situation.

Scenario 1

For each situation, there are **two possible outcomes that Person 1 could observe**. A priori, both outcomes are equally likely. These outcomes are a filled square (left) and a striped square (right).



At the beginning of each situation, **Person 1 will be shown the outcome in that situation**. For example, if the outcome of the situation is a filled square, Person 1 will observe the following:

The outcome in this situation is



In this scenario, **Person 1 is asked to report the observed outcome of the situation to Person 2**, who does not know it. **Person 1's report** will determine the **payoffs** for both Person 1 and Person 2 in each situation. Before deciding what outcome to report, **Person 1 observes all payoffs** associated with each possible report in a table similar to the one presented below.

	Person 1's payoff	Person 2's payoff
If Person 1 reports 	##	##
If Person 1 reports 	##	##

Along with Person 1's message, **Person 2 will only be informed about their own payoff according to Person 1's report**. Hence, Person 2 will not know the observed payoff table corresponding to the message received nor the outcome Person 1 had to report. Moreover, participants in the role of Person 3 have to decide whether to deduct up to 10 points from Person 1's payment for each possible report they can make. **Person 1 is aware of Person 3's role before making their report**. For more details on Person 1's instructions, you can see them in a separate tab by clicking [here](#).

Your task:

You have to **evaluate each possible report Person 1 could make** in each situation. Details about the type of evaluation will be given before showing you the situations. You will make your assessment for each of the two possible reports that Person 1 can make.

Participants in the role of **Person 1 know that other participants in the study will be aware of the possible reports they can make**. However, before making their decisions, Person 1 also knows that no other participant will ever see their actual report in any situation.

After carefully reading all the previous instructions, press the 'Next' button to start your task in Scenario 1.

Next

Figure 59: Instructions of the Reporting setting for observers in the norms part.

Your task: Evaluate based on your own personal opinion.

You have to **evaluate each possible report** that Person 1 could make based on what you consider **personally appropriate** in each situation. Specifically, for **each possible report** that Person 1 could make in each situation, **evaluate according to your own personal opinion** and independently of others' opinion, whether it is appropriate or not to make it. You will rate your personal appropriateness on a scale from "Very personally appropriate" to "Very personally inappropriate". Here, "appropriate" behavior means the one you personally consider to be "morally correct". The standard is, hence, **your personal opinion, independently of the opinion of others**.

Your payment in this task is the **flat rate you are paid for all your evaluations**. There is no right or wrong answer in this part of the study. Please, answer as precisely as possible with your own honest opinion. Your answers in this part will not affect your earnings, and will never be known by any other participant in this study. You can view an example (different from the situations you will evaluate) of how the different actions can be evaluated by clicking [here](#).

After carefully reading the previous instructions, press the 'Next' button to start your task in Scenario 1.

Next

Figure 60: Instructions of the personal norm elicitation for observers.

Situation 1: Evaluate based on your own personal opinion

The **outcome that Person 1 observed** in this situation is:



The table below shows both participants' payoffs corresponding to each **possible report** Person 1 could make.

		Person 1's payoff	Person 2's payoff
If Person 1 reports		10	10
If Person 1 reports		15	5

Please evaluate each possible **report** of Person 1 in this situation according to **your own personal opinion**.

		Very personally appropriate	Personally appropriate	Somewhat personally appropriate	Somewhat personally inappropriate	Personally inappropriate	Very personally inappropriate
If Person 1 reports		☐	☐	☐	☐	☐	☐
If Person 1 reports		☐	☐	☐	☐	☐	☐

Please, press the 'Next' button after completing this page (entering a value for each possible report).

Next

To display the instructions of this task again, click [here](#).

Figure 61: Personal norm elicitation in LG (harmful situation) from the observer's perspective.

Your task: Evaluate based on the opinion of society.

You have to **evaluate each possible report** that Person 1 could make based on what you consider **socially appropriate** in each situation. Specifically, for **each possible report** that Person 1 could make in each situation, **evaluate according to the opinion of society** and independently of your own personal opinion, whether it is appropriate or not to make it. You will rate your social appropriateness on a scale from "Very socially appropriate" to "Very socially inappropriate". Here, "appropriate" behavior means the one you consider most people in your role would agree upon as being "morally correct". The standard is, hence, **not your personal opinion, but your assessment of the opinion of society**. Please try to answer as precisely as possible.

Your payment in this task will **depend on the accuracy of your evaluations**. To determine each participant's payment in Part 2, a computer will **randomly select one single situation and one report**. You will earn an **additional 0.5 GBP (~0.67 USD)** if **your answer coincides with the most frequent evaluation** provided by other participants in your role for that report and situation. Your answers will never be known by any other participant in this study. You can view an example (different from the situations you will evaluate) of how the different actions can be evaluated and how they could be paid by clicking [here](#).

After carefully reading the previous instructions, press the 'Next' button to start your task in Scenario 1.

Next

Figure 62: Instructions of the injunctive norm elicitation for observers.

Situation 1: Evaluate based on the **opinion of society**

The **outcome that Person 1 observed** in this situation is: 

The table below shows both participants' payoffs corresponding to each **possible report** Person 1 could make.

		Person 1's payoff	Person 2's payoff
If Person 1 reports 		10	10
If Person 1 reports 		15	5

Please evaluate each possible **report** of Person 1 in this situation according to **the opinion of society**.

You can earn an **extra 0.5 GBP (~0.67 USD)** if **your answer matches the most frequent evaluation given by others in your role**.

		Very socially appropriate	Socially appropriate	Somewhat socially appropriate	Somewhat socially inappropriate	Socially inappropriate	Very socially inappropriate
If Person 1 reports 		☐	☐	☐	☐	☐	☐
If Person 1 reports 		☐	☐	☐	☐	☐	☐

Please, press the 'Next' button after completing this page (entering a value for each possible report).

Next

To display the instructions of this task again, click [here](#).

Figure 63: Injunctive norm elicitation in LG (harmful situation) from the observer's perspective.

Your task: Estimate participants' reports

You have to **estimate the number** of participants in the role of Person 1 **who made each possible report** in each situation.

Specifically, you have to **estimate the number of participants** in the role of Person 1 (out of 100) **you expect to make each report** in each situation. You will enter your estimate by **clicking a slider and dragging it to your guessed value**. You will observe the numbers you are about to estimate below the slider while dragging it. Please try to answer as precisely as possible.

Your payment in this task will **depend on the accuracy of your evaluations**. To determine each participant's payment in Part 2, a computer will **randomly select one single situation and one report**. You will earn an **additional 0.5 GBP (~0.67 USD)** if **your answer closely matches the actual number** of participants in the role of Person 1 who made that report in that situation (allowing a margin of error of ± 5).

After carefully reading the previous instructions, press the 'Next' button to start your task in Scenario 1.

Next

Figure 64: Instructions of the descriptive norm elicitation for observers.

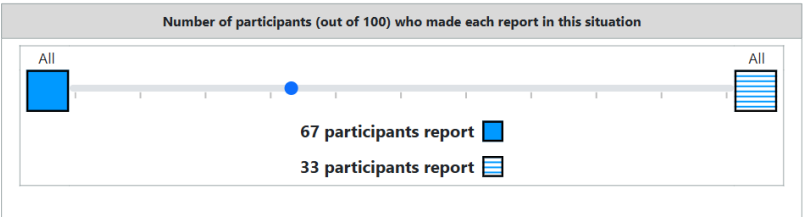
Situation 1: Estimate participants' reports

The outcome that **Person 1 observed** in this situation is: 

The table below shows both participants' payoffs corresponding to each **possible report** Person 1 could make.

	Person 1's payoff	Person 2's payoff
If Person 1 reports 	10	10
If Person 1 reports 	15	5

Please **estimate the number of participants** in the role of Person 1 (out of 100) **who make each possible report in this situation**. You will earn an **additional 0.5 GBP (~0.67 USD)** if **your answer closely matches the actual number** of participants in the role of Person 1 who made the report in this situation (allowing a margin of error of ± 5). Enter your estimation by clicking and dragging the slider below.



Please, press the 'Next' button after completing this page (clicking on the previous slider and dragging it to enter the values).

Next

To display the instructions of this task again, click [here](#).

Figure 65: Descriptive norm elicitation in LG (harmful situation) from the observer's perspective. In this example, the participant guesses that 67/100 senders choose the truth.

C.9.2 Senders' instructions

In this section, we present only the instructions of both settings and an example of the decision pages in each game. Besides, we first present the instructions of the initial page where the experiment is introduced to senders, providing their role and some important details such as the selection of the situation and the payments are explained. As indicated in the sender's procedure (see Figure 50), this page contains an attention check.

Welcome to this study!

The session:

This **session comprises several parts**. For each part, participants in this study will make decisions. These decisions cannot be reversed, and decisions made in one part of the study will not affect the decisions or possible earnings in other parts in the session.

Each participant in this study is assigned the role of either Person 1, Person 2, or Person 3. **You will keep the same role in all parts.** Your role determines which decisions you can make and what payments you can get. Below, you can see your role in the whole study.

Your role:

Person 1

Payments:

Your payment will consist of a **flat rate of 0.64 GBP (~0.86 USD)** and a **bonus payment** based on your decisions and those of other participants. For your bonus, a computer will **randomly select one single part in the session and will randomly match you with other participants**. Decisions in one part will not influence the rest of the parts in the study. Hence, **you should consider each part as if it were the only one selected for your bonus payment**, carefully **reading the instructions for each part** in the study.

We will determine the bonus payment using points. The conversion rate in this study is **10 points = 0.5 GBP (~0.67 USD)**.

You will not be informed of your earnings until the very end of the study. **All your decisions and final earnings will be treated anonymously and confidentially**, that is, we will not disclose your final earnings in the study to any other participant, nor will we identify you with your decisions in the session.

The next lines check if you are reading the page carefully: After reading the entire page, press the number 1 key on your keyboard (either above the letter keys or enabling your numeric keypad) before clicking the 'Next' button at the end of the page to pass this attention check. If you click the 'Next' button directly without previously pressing the number 1 key, you will fail this attention check. Two attention check failures in this study imply an automatic rejection, resulting in no completion code nor payment. Note that this may not be the last attention check you will encounter during this study. These checks help us verify that you are reading the study pages carefully and staying attentive throughout. They are simple and routine, and they do not measure your ability or knowledge. There are no page time limits, but active participation is required throughout the study, so unusually long breaks are not permitted. If a page remains open for an extended period, you may be asked to briefly confirm you are still participating and then continue at your own pace. Non-responses to these confirmations within a short period will result in exclusion.

After carefully reading all the previous instructions, you can proceed with the first part of the study.

Next

Figure 66: Introduction of the experiment for senders

Part 1: Instructions of Scenario 1.

Now, you will take part in some **situations of Scenario 1**, which involves participants in the role of Person 1, 2 and 3. To determine your payments if Part 1 is used, a computer will randomly select one single situation and randomly match you with two other participants. As **Person 1**, you will make decisions in all these situations that might affect your earnings and those of others. Thus, it is **important that you read these instructions**.

Scenario 1

For each situation, there are **two possible outcomes that you could observe**. A priori, both outcomes are equally likely. These outcomes are a filled square (left) and a striped square (right).



At the beginning of each situation, **you will be shown the outcome in that situation**. For example, if the outcome of the situation is a filled square, you will observe the following:

The outcome in this situation is



Your task: In this scenario, we ask you to **report the observed outcome of the situation** to Person 2, who does not know it.

Your report in each situation will **determine the payoffs** that you and Person 2 could earn. For each situation, your payoffs and those of Person 2 will be displayed to you in a table such as the one presented below.

	Your payoff	Person 2's payoff
If you report 	##	##
If you report 	##	##

You will report the outcome of the situation to Person 2 by sending a message of the form:

The outcome in this situation is



For each situation, you will report the outcome by **clicking the corresponding message** you will send to Person 2. Note that, along with your message, **Person 2 will only be informed about their own payoff according to this report**. Hence, Person 2 will not know the observed payoff table corresponding to the message received nor the outcome you had to report.

Person 3 will observe the information you are shown on this page. Their role in this scenario is to decide whether to deduct a number of points (from 0 to 10) from the payoff of Person 1 for each of the two possible reports that Person 1 can make in each situation. Note that Person 3 will not know which outcome you reported. Hence, **no other participant will ever know the veracity of your report**. Your decision does not influence Person 3's earnings, which is a flat rate greater than your final earnings. Note that Person 3's decision will only be implemented for the message you send in the situation randomly selected by the computer to determine the bonus payments.

After carefully reading all the previous instructions, press the 'Next' button to start your task in Scenario 1.

Next

Figure 67: Instructions of the Reporting setting for senders in the Reporting-first condition

The outcome in this round is: 

Please look at the table below. Different **payoffs** are associated with **each outcome you can report** to your matched Person 2 in this round. Note that **Person 2 will never know the payoffs** associated with each outcome.

	Your payoff	Person 2's payoff
If you report: 	10	10
If you report: 	15	5

What **outcome do you want to report** to Person 2 in this round?

Please **report the outcome of this round** by clicking the square of the corresponding message you want to send to Person 2.

In this round, **you send to Person 2** the message:

The outcome in this round is 

The outcome in this round is 

Next

To display the instructions of this task, presented to you at the beginning of this part, click [here](#).

Figure 68: Decision page in the Lying Game (harmful situation) from the sender's perspective when the false message is being selected.

The outcome in this round is: 

Please look at the table below. Different **payoffs** are associated with **each outcome you can report** to your matched Person 2 in this round. Note that **Person 2 will never know the payoffs** associated with each outcome.

	Your payoff	Person 2's payoff
If you report: 	10	10
If you report: 	15	5

What **outcome do you want to report** to Person 2 in this round?

Please **report the outcome of this round** by clicking the square of the corresponding message you want to send to Person 2.

In this round, **you send to Person 2** the message:

The outcome in this round is 

The outcome in this round is 

Next

To display the instructions of this task, presented to you at the beginning of this part, click [here](#).

Figure 69: Decision page in the Truth Game (harmful situation) from the sender's perspective when the true message is being selected.

Part 2: Instructions of Scenario 2.

Now, you will take part in some **situations of a new scenario, Scenario 2**, which involves participants in the role of Person 1, 2 and 3. To determine your payments if Part 2 is used, a computer will randomly select one single situation and randomly match you with two other participants. As **Person 1**, you will make decisions in all these situations that might affect your earnings and those of others. Thus, it is **important that you read these instructions**.

Scenario 2

For each situation, there are **two possible outcomes that you could choose**. These outcomes are a filled circle (left) and a striped circle (right).



Your task: In this scenario, we ask you to **choose the outcome of the situation** and report it to Person 2.

Your choice in each situation will **determine the payoffs** that you and Person 2 could earn. For each situation, your payoffs and those of Person 2 will be displayed to you in a table such as the one presented below.

	Your payoff	Person 2's payoff
If you choose 	##	##
If you choose 	##	##

You will choose the outcome of the situation and report it to Person 2 by sending a message of the form:

I choose



For each situation, you will choose the outcome by **clicking the corresponding message** you will send to Person 2. Note that, along with your message, **Person 2 will only be informed about their own payoff according to this choice**. Hence, Person 2 will not know the observed payoff table corresponding to the message received.

Person 3 will observe the information you are shown on this page. Their role in this scenario is to decide whether to deduct a number of points (from 0 to 10) from the payoff of Person 1 for each of the two possible choices that Person 1 can make in each situation. Note that Person 3 will not know which outcome you chose. Your decision does not influence Person 3's earnings, which is a flat rate greater than your final earnings. Note that Person 3's decision will only be implemented for the message you send in the situation randomly selected by the computer to determine the bonus payments.

After carefully reading all the previous instructions, press the 'Next' button to start your task in Scenario 2.

Next

Figure 70: Instructions of the Choosing game for senders in the Reporting-first condition

Please look at the table below. Different **payoffs** are associated with **each outcome you can choose** in this round. You will report your choice to **Person 2, who will never know the payoffs** associated with each outcome.

	Your payoff	Person 2's payoff
If you choose: 	10	10
If you choose: 	15	5

What **outcome do you want to choose** in this round?

Please **report the outcome you choose** by clicking the circle of the corresponding message you want to send to Person 2.

In this round, **you send to Person 2** the message:

I choose 

I choose 

Next

To display the instructions of this task, presented to you at the beginning of this part, click [here](#).

Figure 71: Decision page in the Choosing Game (harmful situation) from the sender’s perspective. In this example, the harmful choice is being selected.

C.9.3 Receivers’ instructions

As explained at the beginning of the section, here we provide the pages related to the introduction of the study and each part of the experiment, and the pages in the part where the participant is the receiver, both the instructions of the settings and examples of the pages with the message received. The instructions and decision pages of the norm elicitation tasks in each game are those presented in Subsection C.9.1. To facilitate exposition, the receiver decision pages (i.e., the pages in each setting where they get the sender’s message) are presented before displaying the instructions of each setting.

Welcome to this study!

The session:

This **session comprises several parts**. For each part, participants in this study will make decisions. These decisions cannot be reversed, and decisions made in one part of the study will not affect the decisions or possible earnings in other parts in the session.

Payments:

Your payment will consist of a **flat rate of 1.05 GBP (~1.41 USD)** and a **bonus payment** based on your decisions and those of other participants. For your bonus, a computer will **randomly select parts in the session**. Hence, **you should consider each part as if it were the only one selected for your bonus payment**, carefully **reading the instructions for each part** in the study.

We will determine the bonus payment using points. The conversion rate in this study is **10 points = 0.5 GBP (~0.67 USD)**.

You will not be informed of your earnings until the very end of the study. **All your decisions and final earnings will be treated anonymously and confidentially**, that is, we will not disclose your final earnings in the study to any other participant, nor will we identify you with your decisions in the session.

The next lines check if you are reading the page carefully: After reading the entire page, press the number 1 key on your keyboard (either above the letter keys or enabling your numeric keypad) before clicking the 'Next' button at the end of the page to pass this attention check. If you click the 'Next' button directly without previously pressing the number 1 key, you will fail this attention check. Two attention check failures in this study imply an automatic rejection, resulting in no completion code nor payment. Note that this may not be the last attention check you will encounter during this study. These checks help us verify that you are reading the study pages carefully and staying attentive throughout. They are simple and routine, and they do not measure your ability or knowledge. There are no page time limits, but active participation is required throughout the study, so unusually long breaks are not permitted. If a page remains open for an extended period, you may be asked to briefly confirm you are still participating and then continue at your own pace. Non-responses to these confirmations within a short period will result in exclusion.

After carefully reading all the previous instructions, you can proceed with the first part of the study.

Next

Figure 72: Experiment introduction for receivers

Part 1: Instructions for your task in this part.

In this part of the study, you will **observe some scenarios** that involve other randomly selected participants. **For each scenario, your task is to evaluate decisions made in several situations** that might affect other participants' earnings. The evaluations you make in this part of the study will not affect your decisions or possible earnings in any other part and cannot be reversed.

If this part is used to determine your bonus payment, a computer will **randomly select one single situation and will randomly match you with other participants**. Therefore, it is important that you **read the instructions for each scenario and the evaluations you will make carefully**. Your evaluations in one situation will not influence the rest of the situations. Hence, **you should consider each situation as if it were the only one selected for payments**.

Your **payment** in this part will depend on your evaluations. This payment will **not be affected by your decisions or those of others** in any other part of the study. Keep in mind that **you must complete the entire study to receive any payment**.

After reading the prior guidelines, press the 'Next' button to see the first scenario and your task in Part 1 of the study.

Next

Figure 73: Part 1 introduction (norm elicitation) for receivers

Part 2: Instructions for your task in this part.

You have completed Part 1 of this study. Now, you start with Part 2.

In this part of the study, each participant is assigned the role of either Person 1, Person 2, or Person 3. You will be **randomly matched with other participants for the following study part**. Your role determines which decisions you can make and what payments you can get. Below, you can see your role in this part of the study.

Your role:

Person 2

Remember that:

You must complete the entire study to receive the show-up fee and your earnings from previous tasks if Part 1 determines them. Decisions in previous parts do not affect subsequent parts and cannot be reversed. Hence, it is important that you **read the instructions about the features of the next part**: the scenarios' features, the decisions you will make and the decisions others can make before starting your task in each scenario.

After carefully reading all the previous instructions, you can proceed with the second part of the study.

Next

Figure 74: Part 2 introduction (receiver role assigned) for receivers

Message received in Scenario 2

Person 1 sent you the following message:

The outcome in this situation is



Given the message that Person 1 sent you, your earnings in this part are: 10 points.

Please, press the 'Next' button to continue with the study.

Next

Figure 75: Receiver decision page in the reporting setting after getting the sender's message

Message received in Scenario 2

Person 1 sent you the following message:

I choose



Given the message that Person 1 sent you, your earnings in this part are: 5 points.

Please, press the 'Next' button to continue with the study.

Next

Figure 76: Receiver decision page in the Choosing Game after getting the sender's message.

Part 2: Instructions of Scenario 2.

Now, you will participate in a **new scenario, Scenario 2**, which involves participants in the role of Person 1, 2 and 3. Person 1 and Person 3 make decisions in some situations of Scenario 2, but you, as **Person 2**, will not make any decision in this scenario. However, other participants' decisions may affect your earnings and those of others. Thus, it is **important that you read these instructions**.

Scenario 2

For each situation, there are **two possible outcomes that Person 1 could observe**. A priori, both outcomes are equally likely. These outcomes are a filled square (left) and a striped square (right).



At the beginning of each situation, **Person 1 will be shown the outcome in that situation**. For example, if the outcome of the situation is a filled square, Person 1 will see the sentence:

The outcome in this situation is 

In this scenario, **Person 1 is asked to report the observed outcome of the situation to Person 2 (you)**, who does not know it. Different **payoffs** for you and Person 1 will be **associated with each outcome Person 1 can report to you**. For each situation, your payoffs and those of Person 1 will be displayed to Person 1 before making their report in a table such as the one presented below.

	Your payoff	Person 2's payoff
If you report 	##	##
If you report 	##	##

Person 1 will report the outcome of the situation to you by sending a message of the form:

The outcome in this situation is 

Your task: In this scenario, **you will receive Person 1's report**, making no other action apart from receiving their message. Along with Person 1's message, **you will only be informed about your own payoff according to Person 1's report**. Hence, you will not know the observed payoff table corresponding to the message received nor the outcome Person 1 had to report.

Person 3 will observe the information Person 1 is shown on this page. Their role in this scenario is to decide whether to deduct a number of points (from 0 to 10) from the payoff of Person 1 for each of the two possible reports that Person 1 can make in each situation. Note that Person 3 will not know which outcome Person 1 reported. Hence, **no other participant will ever know the veracity of Person 1's report. Participants in the role of Person 1 are aware of Person 3's role** in the scenario. Person 1's decision does not influence Person 3's earnings, which is a flat rate greater than Person 1's final earnings. Note that Person 3's decision will only be implemented for the message that Person 1 send (and you received) in the situation randomly selected by the computer to determine the bonus payments.

After carefully reading all the previous instructions, press the 'Next' button to start your task in Scenario 2.

Next

Figure 77: Instructions of the Reporting setting for receivers

Part 2: Instructions of Scenario 2.

Now, you will participate in a **new scenario**, **Scenario 2**, which involves participants in the role of Person 1, 2 and 3. Person 1 and Person 3 make decisions in some situations of Scenario 2, but you, as **Person 2**, will not make any decision in this scenario. However, other participants' decisions may affect your earnings and those of others. Thus, it is **important that you read these instructions**.

Scenario 2

For each situation, there are **two possible outcomes that Person 1 could choose**. These outcomes are a striped circle (left) and a filled circle (right).



In this scenario, **Person 1 is asked to choose the outcome of the situation and report it to Person 2 (you)**. Different **payoffs** for you and Person 1 will be **associated with each outcome Person 1 can choose**. For each situation, your payoffs and those of Person 1 will be displayed to Person 1 before making their choice in a table such as the one presented below.

	Your payoff	Person 2's payoff
If you choose 	##	##
If you choose 	##	##

Person 1 will choose the outcome of the situation and report it to Person 2 (you) by sending a message of the form:

I choose 

Your task: In this scenario, **you will receive Person 1's choice**, making no other action apart from receiving their message.

Along with Person 1's message, **you will only be informed about your own payoff according to Person 1's choice**. Hence, you will not know the observed payoff table corresponding to the message received.

Person 3 will observe the information Person 1 is shown on this page. Their role in this scenario is to decide whether to deduct a number of points (from 0 to 10) from the payoff of Person 1 for each of the two possible choices that Person 1 can make in each situation. Note that Person 3 will not know which outcome Person 1 chose. **Participants in the role of Person 1 are aware of Person 3's role** in the scenario. Person 1's decision does not influence Person 3's earnings, which is a flat rate greater than Person 1's final earnings. Note that Person 3's decision will only be implemented for the message that Person 1 send (and you received) in the situation randomly selected by the computer to determine the bonus payments.

After carefully reading all the previous instructions, press the 'Next' button to start your task in Scenario 2.

Next

Figure 78: Instructions of the Choosing Game for receivers

Bibliography

- Abeler, Johannes, Anke Becker, and Armin Falk (2014). “Representative evidence on lying costs”. In: *Journal of Public Economics* 113, pp. 96–104.
- Abeler, Johannes, Daniele Nosenzo, and Collin Raymond (2019). “Preferences for Truth-Telling”. In: *Econometrica* 87.4, pp. 1115–1153.
- Albert, Derek A and Daniel Smilek (2023). “Comparing attentional disengagement between Prolific and MTurk samples”. In: *Scientific Reports* 13.1, p. 20574.
- Albertazzi, Andrea (2022). “Individual cheating in the lab: a new measure and external validity”. In: *Theory and Decision* 93.1, pp. 37–67.
- Alempaki, Despoina, Gönül Doğan, and Silvia Saccardo (2019). “Deception and reciprocity”. In: *Experimental Economics* 22.4, pp. 980–1001.
- Alempaki, Despoina, Gönül Doğan, and Yang Yang (2021). “Lying in a foreign language?” In: *Journal of Economic Behavior and Organization* 185, pp. 946–961.
- Alexander, Larry and Michael Moore (2007). “Deontological ethics”. In: *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
- Alger, Ingela and José Ignacio Rivero-Wildemauwe (2024). “Doing the right thing (or not) in a lemons-like situation: on the role of social preferences and Kantian moral concerns”. arXiv:2405.13186.
- Alós-Ferrer, Carlos and Nick Netzer (2010). “The logit-response dynamics”. In: *Games and Economic Behavior* 68.2, pp. 413–427.
- Alós-Ferrer, Carlos and Simon Weidenholzer (2008). “Contagion and efficiency”. In: *Journal of Economic Theory* 143.1, pp. 251–274.
- Andreoni, James, William Harbaugh, and Lise Vesterlund (2003). “The carrot or the stick: Rewards, punishments, and cooperation”. In: *American Economic Review* 93.3, pp. 893–902.

- Andrighetto, Giulia et al. (2013). "Punish and Voice: Punishment Enhances Cooperation when Combined with Norm-Signalling". In: *PLoS ONE* 8.6, pp. 1–8.
- Axelrod, Robert (1986). "An evolutionary approach to norms". In: *American Political Science Review* 80.4, pp. 1095–1111.
- Aycinena, Diego et al. (2022). "Social norms and dishonesty across societies". In: *Proceedings of the National Academy of Sciences of the United States of America* 119.31.
- Babin, J. Jobu and Haritima S. Chauhan (2024). "Replication: Erat and Gneezy's white lies paradigm". In: *Journal of Economic Psychology* 102, p. 102704.
- Barr, Abigail, Tom Lane, and Daniele Nosenzo (2018). "On the social inappropriateness of discrimination". In: *Journal of Public Economics* 164, pp. 153–164.
- Bašić, Zvonimir and Simone Quercia (2022). "The Influence of Self and Social Image Concerns on Lying". In: *Games and Economic Behavior* 133, pp. 162–169.
- Bašić, Zvonimir and Eugenio Verrina (2024). "Personal norms — and not only social norms — shape economic behavior". In: *Journal of Public Economics* 239, p. 105255.
- Behnk, Sascha, Iván Barreda-Tarrazona, and Aurora García-Gallego (2018). "Punishing liars-How monitoring affects honesty and trust". In: *PLoS ONE* 13.10, pp. 1–30.
- Behnk, Sascha, Li Hao, and Ernesto Reuben (2022). "Shifting normative beliefs: On why groups behave more antisocially than individuals". In: *European Economic Review* 145, p. 104116.
- Bénabou, Roland, Armin Falk, and Luca Henkel (2024). *Ends versus Means: Kantians, Utilitarians, and Moral Decisions*. Working Paper 32073. National Bureau of Economic Research.
- Bénabou, Roland, Armin Falk, Luca Henkel, and Jean Tirole (2020). "Eliciting Moral Preferences: Theory and Experiment". Working Paper.
- Berg, Joyce, John Dickhaut, and Kevin McCabe (1995). "Trust, reciprocity, and social history". In: *Games and economic behavior* 10.1, pp. 122–142.
- Bergin, James and Barton L Lipman (1996). "Evolution with state-dependent mutations". In: *Econometrica* 64.4, pp. 943–956.
- Besancenot, Damien and Radu Vranceanu (2020). "Profession and deception: Experimental evidence on lying behavior among business and medical students". In: *Journal of Economic Behavior & Organization* 179, pp. 175–187.

- Bicchieri, Cristina (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, pp. 1–260.
- Bicchieri, Cristina, Eugen Dimant, Simon Gächter, et al. (2022). “Social proximity and the erosion of norm compliance”. In: *Games and Economic Behavior* 132, pp. 59–72.
- Bicchieri, Cristina, Eugen Dimant, and Silvia Sonderegger (2023). “It’s Not a Lie If You Believe the Norm Does Not Apply: Conditional Norm-Following with Strategic Beliefs”. In: *Games and Economic Behavior* 138, pp. 321–354.
- Bicchieri, Cristina, Eugen Dimant, and Erte Xiao (2021). “Deviant or wrong? The effects of norm information on the efficacy of punishment”. In: *Journal of Economic Behavior and Organization* 188, pp. 209–235.
- Bicchieri, Cristina and Marta Maras (2022). “Intentionality matters for third-party punishment but not compensation in trust games”. In: *Journal of Economic Behavior and Organization* 197, pp. 205–220.
- Bilancini, Ennio, Juan Francisco Blazquiz Pulido, and Leonardo Boncinelli (2023). “Evolution of Conventions in Uncertain Environments”. In: *Available at SSRN 4360114*. Working paper.
- Bilancini, Ennio and Leonardo Boncinelli (2020). “The evolution of conventions under condition-dependent mistakes”. In: *Economic Theory* 69.2, pp. 497–521.
- Bilancini, Ennio, Leonardo Boncinelli, and Heinrich H Nax (2021). “What noise matters? Experimental evidence for stochastic deviations in social norms”. In: *Journal of Behavioral and Experimental Economics* 90, p. 101626.
- Biziou-Van-Pol, Laura et al. (2015). “Does telling white lies signal pro-social preferences?” In: *Judgment and Decision Making* 10.6, pp. 538–548.
- Blazquiz-Pulido, Juan Francisco (2026). “Shooting the Messenger? The Role of Consequences and (Dis)Honesty in Punishment”. Working paper.
- Blazquiz-Pulido, Juan Francisco, Luca Polonio, and Ennio Bilancini (2024). “Who’s the deceiver? Identifying deceptive intentions in communication”. In: *Games and Economic Behavior* 145, pp. 451–466.
- Blume, Lawrence E (1993). “The statistical mechanics of strategic interaction”. In: *Games and Economic Behavior* 5.3, pp. 387–424.
- (2003). “How noise matters”. In: *Games and Economic Behavior* 44.2, pp. 251–271.

- Brandts, Jordi and Gary Charness (2003). "Truth or consequences: An experiment". In: *Management Science* 49.1, pp. 116–130.
- Capraro, Valerio (2017). "Does the truth come naturally? Time pressure increases honesty in one-shot deception games". In: *Economics Letters* 158, pp. 54–57.
- (2018). "Gender differences in lying in sender-receiver games: A meta-analysis". In: *Judgment and Decision Making* 13.4, pp. 345–355.
- Capraro, Valerio, Joseph Y Halpern, and Matjaž Perc (2024). "From outcome-based to language-based preferences". In: *Journal of Economic Literature* 62.1, pp. 115–54.
- Capraro, Valerio and Matjaž Perc (2021). "Mathematical foundations of moral preferences". In: *Journal of the Royal Society Interface* 18.175.
- Capraro, Valerio, Jonathan Schulz, and David G Rand (2019). "Time pressure and honesty in a deception game". In: *Journal of Behavioral and Experimental Economics* 79, pp. 93–99.
- Charness, Gary and Martin Dufwenberg (2006). "Promises and partnership". In: *Econometrica* 74.6, pp. 1579–1601.
- Chavez, Alex K. and Cristina Bicchieri (2013). "Third-party sanctioning and compensation behavior: Findings from the ultimatum game". In: *Journal of Economic Psychology* 39, pp. 268–277.
- Cheng, Xuemei et al. (2022). "Punishment cost affects third-parties' behavioral and neural responses to unfairness". In: *International Journal of Psychophysiology* 177. March, pp. 27–33.
- Civai, Claudia, Valerio Capraro, and Luca Polonio (2025). "The role of attention and frames on third-party punishment and compensation choices". In: *Cognition* 263, p. 106192.
- D'Adda, Giovanna, Michalis Drouvelis, and Daniele Nosenzo (2016). "Norm elicitation in within-subject designs: Testing for order effects". In: *Journal of Behavioral and Experimental Economics* 62, pp. 1–7.
- D'Adda, Giovanna, Martin Dufwenberg, et al. (2020). "Social norms with private values: Theory and experiments". In: *Games and Economic Behavior* 124.2020, pp. 288–304.
- Dickinson, David L., David Masclet, and Marie Claire Villeval (2015). "Norm enforcement in social dilemmas: An experiment with police commissioners". In: *Journal of Public Economics* 126, pp. 74–85.
- Dickinson, David L. and David M McEvoy (2021). "Further from the truth: The impact of moving from in-person to online settings on dishonest behavior". In: *Journal of Behavioral and Experimental Economics* 90, p. 101649.

- Dimant, Eugen, Fabio Galeotti, and Marie Claire Villeval (2024). "Motivated Information Acquisition and Social Norm Formation". In: *European Economic Review* 167, p. 104778.
- Dimant, Eugen, Michele Gelfand, et al. (2025). "Strategic behavior with tight, loose, and polarized norms". In: *Management Science* 71.3, pp. 2245–2263.
- Dimant, Eugen and Tobias Gesche (2023). "Nudging enforcers: How norm perceptions and motives for lying shape sanctions". In: *PNAS Nexus* 2.7, pgad224.
- Douglas, Benjamin D, Patrick J Ewell, and Markus Brauer (2023). "Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA". In: *PLoS ONE* 18.3, e0279720.
- Eisenkopf, Gerald, Ruslan Gurtoviy, and Verena Utikal (2017). "Punishment Motives for Small and Big Lies". In: *Journal of Economics and Management Strategy* 26.2, pp. 484–498.
- Ellison, Glenn (2000). "Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution". In: *Review of Economic Studies* 67.1, pp. 17–45.
- Ely, Jeffrey C (2002). "Local conventions". In: *Advances in Theoretical Economics* 2.1.
- Erat, Sanjiv and Uri Gneezy (2012). "White lies". In: *Management Science* 58.4, pp. 723–733.
- Erkut, Hande, Daniele Nosenzo, and Martin Sefton (2015). "Identifying social norms using coordination games: Spectators vs. stakeholders". In: *Economics Letters* 130, pp. 28–31.
- Evans, R (1993). "Observability, Imitation and Cooperation in the Repeated Prisoners' Dilemma". In: *Unpublished, Cambridge University*.
- Feess, Eberhard, Florian Kerzenmacher, and Yuriy Timofeyev (2022). "Utilitarian or deontological models of moral behavior—what predicts morally questionable decisions?" In: *European Economic Review* 149, p. 104264.
- Fehr, Ernst and Urs Fischbacher (2004). "Third-party punishment and social norms". In: *Evolution and Human Behavior* 25.2, pp. 63–87.
- Fehr, Ernst and Bettina Rockenbach (2003). "Detrimental effects of sanctions on human altruism". In: *Nature* 422.6928, pp. 137–140.
- Fischbacher, Urs and Franziska Föllmi-Heusi (2013). "Lies in disguise—an experimental study on cheating". In: *Journal of the European Economic Association* 11.3, pp. 525–547.

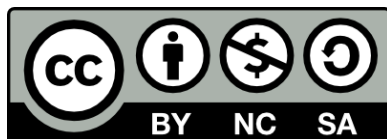
- Foster, Dean and Peyton Young (1990). "Stochastic evolutionary game dynamics". In: *Theoretical Population Biology* 38.2, pp. 219–232.
- Freidlin, Mark Iosifovich and Alexander D Wentzell (1984). *Random Perturbations of Dynamical Systems*. Springer.
- Gächter, Simon, Elke Renner, and Martin Sefton (2008). "The long-run benefits of punishment". In: *Science* 322.5907, pp. 1510–1510.
- Germano, Fabrizio (2007). "Stochastic evolution of rules for playing finite normal form games". In: *Theory and Decision* 62, pp. 311–333.
- Glätzle-Rützler, Daniela and Philipp Lergetporer (2015). "Lying and age: An experimental study". In: *Journal of Economic Psychology* 46, pp. 12–25.
- Gneezy, Uri (2005). "Deception: The role of consequences". In: *American Economic Review* 95.1, pp. 384–394.
- Gneezy, Uri, Agne Kajackaite, and Joel Sobel (2018). "Lying Aversion and the Size of the Lie". In: *American Economic Review* 108.2, pp. 419–453.
- Gneezy, Uri, Bettina Rockenbach, and Marta Serra-Garcia (2013). "Measuring lying aversion". In: *Journal of Economic Behavior and Organization* 93, pp. 293–300.
- Grundherr, Michael von, Johanna Jauernig, and Matthias Uhl (2021). "To condemn is not to punish: An experiment on hypocrisy". In: *Games* 12.2, p. 38.
- Gupta, Neeraja, Luca Rigotti, and Alistair Wilson (2021). "The experimenters' dilemma: inferential preferences over populations". In: *arXiv preprint arXiv:2107.05064*.
- Gylfason, Haukur Freyr and Katrin Olafsdottir (2017). "Does Gneezy's cheap talk game measure trust?" In: *Journal of Behavioral and Experimental Economics* 67, pp. 143–148.
- Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian Journal of Statistics* 6, pp. 65–70.
- Huber, Christoph et al. (2023). "On social norms and observability in (dis)honest behavior". In: *Journal of Economic Behavior and Organization* 121, pp. 1086–1099.
- Hurkens, Sjaak and Navin Kartik (2009). "Would i lie to you? on social preferences and lying aversion". In: *Experimental Economics* 12.2, pp. 180–192.
- Hwang, Sung-Ha, Wooyoung Lim, et al. (2018). "Conventional contracts, intentional behavior and logit choice: Equality without symmetry". In: *Games and Economic Behavior* 110, pp. 273–294.

- Hwang, Sung-Ha, Suresh Naidu, and Samuel Bowles (2016). *Social conflict and the evolution of unequal conventions*. Working Paper. Santa Fe Institute.
- Ianni, Antonella (2014). "Learning strict Nash equilibria through reinforcement". In: *Journal of Mathematical Economics* 50, pp. 148–155.
- Janezic, Katharina A (2020). "Heterogeneity in lies and lying preferences". Working paper.
- Jordan, Jillian J., Katherine McAuliffe, and David Rand (2016). "The effects of endowment size and strategy method on third party punishment". In: *Experimental Economics* 19.4, pp. 741–763.
- Jordan, Jillian J. and David G Rand (2020). "Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions." In: *Journal of Personality and Social Psychology* 118.1, p. 57.
- Kajackaite, Agne and Uri Gneezy (2017). "Incentives and cheating". In: *Games and Economic Behavior* 102, pp. 433–444.
- Kandori, Michihiro, George J Mailath, and Rafael Rob (1993). "Learning, mutation, and long run equilibria in games". In: *Econometrica* 61.1, pp. 29–56.
- Kant, Immanuel (1785). *Grundlegung zur Metaphysik der Sitten*. Johann Friedrich Hartknoch.
- (1797). "On a supposed right to lie from philanthropy". In: *Practical philosophy* 612.
- Kawagoe, Toshiji and Hirokazu Takizawa (2009). "Equilibrium refinement versus level-k analysis: An experimental study of cheap-talk games with private information". In: *Games and Economic Behavior* 66.1, pp. 238–255.
- Khalmetski, Kiryl, Bettina Rockenbach, and Peter Werner (2017). "Evasive lying in strategic communication". In: *Journal of Public Economics* 156, pp. 59–72.
- Khalmetski, Kiryl and Dirk Sliwka (2019). "Disguising lies - Image Concerns and Partial Lying in Cheating Games". In: *American Economic Journal: Microeconomics* 11.4, pp. 79–110.
- Kim, Chang-Jin (1994). "Dynamic linear models with Markov-switching". In: *Journal of Econometrics* 60.1-2, pp. 1–22.
- Kimbrough, Erik O. and Alexander Vostroknutov (2016). "Norms Make Preferences Social". In: *Journal of the European Economic Association* 14.3, pp. 608–638.
- (2023). "Resentment and Punishment". In: *SSRN Electronic Journal*.

- Krupka, Erin L. and Roberto A Weber (2013). "Identifying social norms using coordination games: Why does dictator game sharing vary?" In: *Journal of the European Economic Association* 11.3, pp. 495–524.
- Laske, Katharina, Silvia Saccardo, and Uri Gneezy (2018). "Do Fines Deter Unethical Behavior? The Effect of Systematically Varying the Size and Probability of Punishment". In: *SSRN Electronic Journal*.
- Leibbrandt, Andreas and Raúl López-Pérez (2012). "An exploration of third and second party punishment in ten simple games". In: *Journal of Economic Behavior and Organization* 84.3, pp. 753–766.
- Lewis, David (2008). *Convention: A philosophical study*. John Wiley & Sons.
- Li, Xiaolin, Özalp Özer, and Upender Subramanian (2022). "Are we strategically naïve or guided by trust and trustworthiness in cheap-talk communication?" In: *Management Science* 68.1, pp. 376–398.
- Lim, Wooyoung and Philip R Neary (2016). "An experimental investigation of stochastic adjustment dynamics". In: *Games and Economic Behavior* 100, pp. 208–219.
- Lohse, Tim, Sven A. Simon, and Kai A. Konrad (2018). "Deception under time pressure: Conscious decision or a problem of awareness?" In: *Journal of Economic Behavior and Organization* 146, pp. 31–42.
- López-Pérez, Raúl and Eli Spiegelman (2013). "Why do people tell the truth? Experimental evidence for pure lie aversion". In: *Experimental Economics* 16.3, pp. 233–247.
- (2019). "Do economists lie more?" In: *Dishonesty in behavioral economics*. Academic Press, pp. 143–162.
- Mäs, Michael and Heinrich H Nax (2016). "A behavioral study of "noise" in coordination games". In: *Journal of Economic Theory* 162, pp. 195–208.
- McEvily, Bill, Joseph R Radzevick, and Roberto A Weber (2012). "Whom do you distrust and how much does it cost? An experiment on the measurement of trust". In: *Games and Economic Behavior* 74.1, pp. 285–298.
- Naidu, Suresh, Sung-Ha Hwang, and Samuel Bowles (2010). "Evolutionary bargaining with intentional idiosyncratic play". In: *Economics Letters* 109.1, pp. 31–33.
- Nax, Heinrich H and Jonathan Newton (2019). "Risk attitudes and risk dominance in the long run". In: *Games and Economic Behavior* 116, pp. 179–184.
- Newton, Jonathan (2018). "Evolutionary game theory: a renaissance". In: *Games* 9.2, p. 31.
- (2021). "Conventions under heterogeneous behavioural rules". In: *The Review of Economic Studies* 88.4, pp. 2094–2118.

- Nosenzo, Daniele, Erte Xiao, and Nina Xue (2024). "The motive matters: Experimental evidence on the expressive function of punishment". In: *Games and Economic Behavior* 148, pp. 44–67.
- Palan, Stefan and Christian Schitter (2018). "Prolific.ac—A subject pool for online experiments". In: *Journal of Behavioral and Experimental Finance* 17, pp. 22–27.
- Panizza, Folco et al. (2023). "Measuring Norm Pluralism and Tolerance". In: *SSRN Electronic Journal*.
- Peer, Eyal, Laura Brandimarte, et al. (2017). "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research". In: *Journal of experimental social psychology* 70, pp. 153–163.
- Peer, Eyal, David Rothschild, et al. (2022). "Data quality of platforms and panels for online behavioral research". In: *Behavior research methods* 54.4, pp. 1643–1662.
- Peeters, Ronald, Marc Vorsatz, and Markus Walzl (2013). "Truth, Trust, and Sanctions: On Institutional Selection in Sender-Receiver Games". In: *Scandinavian Journal of Economics* 115.2, pp. 508–548.
- (2015). "Beliefs and truth-telling: A laboratory experiment". In: *Journal of Economic Behavior and Organization* 113, pp. 1–12.
- Riehm, Tobias et al. (2022). "Social norms, sanctions, and conditional entry in markets with externalities: Evidence from an artefactual field experiment". In: *Journal of Public Economics* 212, p. 104701.
- Robson, Arthur J and Fernando Vega-Redondo (1996). "Efficient equilibrium selection in evolutionary games with random matching". In: *Journal of Economic Theory* 70.1, pp. 65–92.
- Sánchez-Pagés, Santiago and Marc Vorsatz (2007). "An experimental study of truth-telling in a sender-receiver game". In: *Games and Economic Behavior* 61.1, pp. 86–112.
- (2009). "Enjoy the silence: An experiment on truth-telling". In: *Experimental Economics* 12.2, pp. 220–241.
- Sandholm, William H (2010). *Population games and evolutionary dynamics*. MIT press.
- Sandholm, William H, Segismundo S Izquierdo, and Luis R Izquierdo (2019). "Best experienced payoff dynamics and cooperation in the centipede game". In: *Theoretical Economics* 14.4, pp. 1347–1385.
- (2020). "Stability for best experienced payoff dynamics". In: *Journal of Economic Theory* 185, p. 104957.
- Sawa, Ryoji and Jiabin Wu (2018). "Prospect dynamics and loss dominance". In: *Games and Economic Behavior* 112, pp. 98–124.

- Schudy, Simeon, Susanna Grundmann, and Lisa Spantig (2025). "Individual preferences for truth-telling". CESifo Working Paper No. 11521.
- Sobel, Joel (2020). "Lying and deception in games". In: *Journal of Political Economy* 128.3, pp. 907–947.
- Sun, Xuezheng and Zhao Yang (2008). "Generalized McNemar's test for homogeneity of the marginal distributions". In: *SAS Global forum*. Vol. 382, pp. 1–10.
- Suri, Dominik, Sebastian Kube, and Johannes Schultz (2025). "Evaluating online data collection platforms using a simple rule-following task". In: *Economics Letters* 255, p. 112509.
- Sutter, Matthias (2009). "Deception through telling the truth?! Experimental evidence from individuals and teams". In: *Economic Journal* 119.534, pp. 47–60.
- Thielmann, Isabel, Giuliana Spadaro, and Daniel Balliet (2020). "Personality and prosocial behavior: A theoretical framework and meta-analysis." In: *Psychological Bulletin* 146.1, p. 30.
- Van Leeuwen, Boris and Ingela Alger (2024). "Estimating social preferences and Kantian morality in strategic interactions". In: *Journal of Political Economy Microeconomics* 2.4, pp. 665–706.
- Vanberg, Christoph (2017). "Who never tells a lie?" In: *Experimental Economics* 20.2, pp. 448–459.
- Wang, Joseph Tao Yi, Michael Spezio, and Colin F. Camerer (2010). "Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games". In: *American Economic Review* 100.3, pp. 984–1007.
- Xiao, Erte (2013). "Profit-seeking punishment corrupts norm obedience". In: *Games and Economic Behavior* 77.1, pp. 321–344.
- (2018). "Punishment, social norms, and cooperation". In: *Research handbook on behavioral law and economics*. Edward Elgar Publishing, pp. 155–173.
- Young, H Peyton (1993). "The evolution of conventions". In: *Econometrica* 61, pp. 57–84.
- (1996). "The economics of convention". In: *Journal of Economic Perspectives* 10.2, pp. 105–122.
- (1998). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.
- (2015). "The evolution of social norms". In: *Annual Review of Economics* 7.1, pp. 359–387.



Unless otherwise expressly stated, all original material of whatever nature created by Juan Francisco Blázquez Pulido and included in this thesis, is licensed under a Creative Commons Attribution Non-commercial Share Alike 3.0 Italy License.

Check on Creative Commons site:

<https://creativecommons.org/licenses/by-nc-sa/3.0/it/legalcode/>

<https://creativecommons.org/licenses/by-nc-sa/3.0/it/deed.en>

Ask the author about other uses.