

# Sex, Lies, and Punishment: Gender Punishment Gap After Suspected Dishonesty

Pia Pico\*, Rafael Teixeira†

November 3, 2023

## Abstract

The punishment gap refers to minorities and underprivileged groups facing more severe punishment for mistakes or transgressions compared to other groups. Untangling the reasons behind the punishment gap is challenging due to the complex interplay of various factors. In an effort to investigate such situations, we analyzed gender differences in receiving punishment for suspected dishonesty within a controlled experiment, using a sender-receiver game. Our results show that males engage in selfish lies more frequently than females and are more likely to be punished for such lies. We also explored individuals' beliefs about dishonesty, considering empirical expectations, normative expectations, and gender-based causal attribution as potential explanations for the observed behaviors. To capture this causal attribution, we developed a new methodology to incentivize the investigation of how people perceive the same behavior differently when it is done by males or females. The findings suggest that empirical expectations might explain the observed behavior, as males are expected to lie more often. The results also suggest that participants perceive lying behaviors to be less socially acceptable than honestly disclosing a selfish outcome. Additionally, lies are often attributed to motives involving 'rational calculations,' with females showing higher levels of such attributions compared to males.

## 1 Introduction

Discrimination against certain groups of people can take many forms. For example, women may be directly excluded from the job application process because of an existing pregnancy, or a person may be denied gym membership because of their ethnicity. But discrimination can also be more subtle. Consequences that individuals experience for misconduct and unethical behavior may vary depending on their membership in certain groups. This so-called punishment gap has already been identified in different contexts. Empirical investigations have shown a punishment gap between men and women for unethical behavior. For example, women working in the financial advisory industry were found to be more often punished than men for unethical behavior in the work setting (Egan, Matvos, & Seru, 2022). Also in the medical field, it was found that female surgeons get punished for a negative outcome attributable to their own but also to other female surgeons' conduct, while this is not the case for male surgeons (Sarsons, 2017).

---

\*University of Kassel

†University of Amsterdam

In this paper, we aim to answer whether those differences in punishment can be attributed to divergent norms regarding ethical behavior for men and women. And if so, whether they translate into gender discrimination in punishing unethical behavior. Since previous findings regarding a punishment gap for unethical behavior between men and women rely on real-world data we aim to establish a clear causal relationship between an individual's gender and the punishment received after suspected unethical behavior and identify in that context the underlying role of social expectations. Moreover, examining how gender influences the punishment of suspected unethical behavior and how gender roles may account for this influence is crucial in advancing gender equality, which is one of the Sustainable Development Goals outlined in the Agenda 2030, part of the collective goals, priorities, and initiatives that the United Nations as an organization pursues. In order to investigate the punishment gap, specifically across gender, we conducted an online experiment in which participants play a sender-receiver game where we incorporated the option to lie as a way to measure unethical behavior and the option to punish suspected unethical behavior. By conducting an experiment in which we only manipulate the information about the senders' gender that is given to the receiver, we are able to rule out other aspects that may lead to gender differences. Additionally, we gather individuals' empirical expectations regarding the prevalence of lying in each gender, as well as the normative expectation associated with lying behavior for each gender. These measures aim to analyze how social norms (Bicchieri (2016)) can impact any potential punishment gap.

To further understand gender differences and the underlying motivations behind the received punishment, our study introduces a novel incentivized tool aimed at measuring perceived causal attributions for each gender. Existing research highlights that males and females are perceived differently despite engaging in similar behaviors. For instance, Smith, Rosenstein, Nikolov, and Chaney (2019) reveals that individuals assess male and female army officers disparately, associating males with analytical or arrogant traits, and females with compassion and temperament. Moreover, individuals' perceptions of an individual and their motives can significantly impact the levels of punishment administered (Sommers and Ellsworth (2000)). We study motivations behind punishment using variants of Krupka and Weber (2013)'s method by adapting their approach to develop a new incentivized measure, allowing us to examine causal attributions and potential motivations, such as "rational calculations" and "emotional decisions", regarding lies for each gender.

The findings of our study confirm prior research on lying behavior, suggesting that men are more likely than women to engage in unethical behavior and perform selfish lies. Our results show the importance of uncertainty associated with a decrease in confidence, as the prevalence of punishment for men increases with the number of rounds played. Additionally, the results align with empirical expectations that men are more likely to engage in lying behavior. Lying behavior is also associated with lower levels of appropriateness compared to truly revealing the selfish outcome. Participants are more likely to attribute lies to motivations such as 'rational calculations' and 'malicious intentions' instead of mistakes. While no major gender difference is observed for normative expectations and attributions, lying females are considered more rational than lying males.

The paper is structured as follows: Section 2 presents the literature review and outlines our hypothesis, Section 3 discusses the experiment's design, Section 4 presents the results, Section 5 explores the implications of the findings, and Section 6 provides the conclusion.

## 2 Literature Review

This section is divided into four distinct parts: the first part provides an overview of the literature and hypotheses related to gender differences in lying behavior. Following that, the second part discusses the literature and hypotheses surrounding punishment and the punishment gap. The third part explores the literature and hypotheses that connect the preceding subsections with social norms and beliefs. Finally, the fourth part introduces a novel method for capturing attribution and elucidates its relationship with the potential punishment gap and gender differences.

### 2.1 Gender differences in lying behavior

A way to measure unethical behavior is to give people the choice to lie (Kouchaki and Smith (2014); Laske, Saccardo, and Gneezy (2018)) about an uncontrollable outcome that defines their monetary outcome. Those lies may result in an advantageous outcome for themselves which, however, negatively affects the outcome for another person, or in contrast, lies can lead to a disadvantageous outcome for themselves leading to a positive outcome for someone else.

Individuals' lying behavior has been widely studied in the last 15 years. Among other things, gender difference in lying behavior was investigated in depth (e.g. Dreber and Johannesson (2008); Friesen and Gangadharan (2012); Erat and Gneezy (2012); Grosch and Rau (2017); Capraro and Peltola (2018); Cappelen, Konow, Sørensen, and Tungodden (2013); Biziou-van Pol, Haenen, Novaro, Liberman, and Capraro (2015)). Though research on gender differences in lying behavior is manifold, a clear direction of which gender is more likely to lie could not be established. One argument is that the mixed results regarding the role of gender come from the different consequences lying can have for the individual itself and for others affected by the lie (Capraro & Peltola, 2018).

Most closely related to our research, using a meta-analysis with over 8700 observations, Capraro and Peltola (2018) find that men are significantly more likely to tell selfish and prosocial lies than women, showing that independently of the direct consequences of the lie, men engage more in lying behavior than women. Since most studies find a difference in telling selfish and prosocial lies between men and women, our first goal is to replicate those results by testing our first hypotheses.

**H1.1.** There are gender differences in telling prosocial lies.

**H1.2.** There are gender differences in telling selfish lies.

### 2.2 Gender differences in punishment

Punishment has been widely studied for more than two decades. Here, one can differentiate between receiving punishment and giving punishment. Gender differences in receiving and giving punishment are studied using, among other methods, Trust games (Croson and Buchan (1999); Dittrich (2015)), Ultimatum games (Saad and Gill (2001); García-Gallego, Georgantzis, and Jaramillo-Gutiérrez (2012)), Public Goods games (Burnham, 2018), Punishment games (Eckel & Grossman, 1996) and Corruption games (Fišar, Kubák, Špalek, & Tremewan, 2016), whereas more literature exists that focuses at giving punishment than on receiving punishment.

For giving and receiving punishment, existing literature is inconclusive since some studies find that men punish more (Burnham (2018); Croson and Buchan (1999); Fišar et al. (2016)) and are punished more (Saad & Gill, 2001) for antisocial behavior than females are, others find that females are more likely to punish (Eckel & Grossman, 1996) and being punished for antisocial behavior (Mieth, Buchner, and Bell (2017); Jung and Vranceanu (2015)). These differences in results may be attributed to their different experimental approaches.

Since the literature on gender differences in receiving punishment is not as extensive as the literature on gender differences in giving punishment, we are particularly interested in observing potential gender differences in receiving punishment.

Eisenkopf, Gurtoviy, and Utikal (2017) conduct a laboratory experiment to disentangle individuals' reactions to dishonest behavior which results in more or less economic harm. They find that the severity of the lie leads to an increase in punishment. Another study by Peeters, Vorsatz, and Walzl (2013) implements a sender-receiver game in which the sender can lie about two different payoff divisions between the sender and receiver that are equally likely to be drawn by nature and that result in either a higher payoff for the sender or for the receiver. After the sender gives the information to the sender, the receiver is informed about the truth and is given the option to punish the sender which results in zero payoffs for both players. The results show that receivers forgo private payoffs in order to punish senders for lies. Similar results were found by Sánchez-Pagés and Vorsatz (2007a) and Sánchez-Pagés and Vorsatz (2009). Although lying shown to be a powerful incentive to engage in costly punishment in general, none of the studies explore gender differences in this context. Therefore, our second hypothesis is testing gender differences in receiving punishment for selfish and prosocial lies.

**H2.1.** There are gender differences in receiving punishment for telling prosocial lies.

**H2.2.** There are gender differences in receiving punishment for telling selfish lies.

### 2.3 Social norms, punishment and discrimination

Social norms play a pivotal role in our research, serving two key purposes. Initially, we delve into the connection between social norms and punishment and its relations with gender, shedding light on the reasons behind punitive actions. Subsequently, we describe how social norms intersect with discrimination, potentially giving rise to a punishment gap.

We initiate our exploration by analyzing research that delves into the interplay between social norms and the consequences of unethical behavior. Additionally, we closely examine how different genders experience different impacts when they transgress these norms. Social norms can be expressed in different terms, but we use the terminology developed by (Bicchieri, 2016) who divides norms into empirical expectations (beliefs about what we expect others to do) and normative expectations (beliefs about what others think we should do). Lying, if it is to the disadvantage of the individuals not telling it but being affected by it, can be seen as unethical behavior (Brandts & Charness, 2003). Dictator games have shown that women are more altruistic than men (Brañas-Garza, Capraro, and Rascon-Ramirez (2018); Falk and Hermle (2018)) and that women are expected to be more altruistic which is a shared opinion of males and females (Brañas-Garza et al., 2018). However, it was also shown that women are only more altruistic than men when participants were reminded of their gender (Visser & Roelofs, 2011) which goes along with the findings of Brañas-Garza et al. (2018) and indicates that women are aware of the gender-specific norm of being altruistic.

tic. Heilman and Chen (2005) find experimental evidence that women are evaluated worse than men for not engaging in altruistic behavior in a work setting. This shows that when women are not behaving as they are expected to, they will face negative consequences which corresponds to the argumentation of Cialdini and Trost (1998) who state that violations of normative role prescriptions are penalized. Further, punishment closely orientates at norm perception, leading to a positive relationship between the size of punishment and the severity of the norm violation (Dimant & Gesche, 2021). Since we predict gender differences in normative and empirical expectations of individuals' lying behavior, we believe that a violation against those normative expectations translates into differences in the receipt of punishment, and therefore, shapes the expectation of receiving punishment.

Similarly, these explanations are closely related to the literature on discrimination. For example, the theory developed by Becker (2010), which has been applied in various discussions such as Neumark (2018); Edelman, Luca, and Svirsky (2017), describes that discrimination can originate from statistical reasoning (involving beliefs about differing behavior among groups) or taste-based preferences (favoring one group over another). Empirical expectations can partially capture the statistical reasoning associated with discrimination, while normative expectations (and the causal attribution discussed next) can partially capture a sense of taste-based discrimination. If a particular gender is expected to lie more often, this theory suggests that this gender would be more heavily punished, even if it does not reflect actual behavior. Meanwhile, if people consider a certain gender as less socially acceptable to lie, they could face more punishment, as there is a normative feeling that this gender should not behave in such a manner.

**H3.1.** There are gender differences in empirical expectations regarding engaging in selfish and prosocial lies.

**H3.2.** There are gender differences in normative expectations regarding engaging in selfish lies and truthfully reporting selfish outcomes.

## 2.4 'Social' Attribution

Attribution theory (e.g., Ryan and Connell (1989); Heider (2013); Shaver (2016)) explores how individuals perceive the causes and motivations behind everyday experiences, constructing possible explanations through social inferences based on the context and the individuals involved. This theory has been linked to the punishment gap (e.g. Sommers and Ellsworth (2000)), which describes that, for example, black defendants are considered more guilty, aggressive, and violent than white defendants, resulting in more severe punishment for the former.

Likewise, several studies consistently reveal that males and females can be perceived differently, even when exhibiting the same behavior. For instance, males tend to attribute success in exams to their ability, while females are more likely to attribute the same success to their effort (Beyer (1998)). Additionally, male leaders are generally attributed to be more competent (Garcia-Retamero and Lopez-Zafra (2009)), and female leaders expressing anger are often considered "emotional" and "out of control" compared to their male counterparts (Brescoll and Uhlmann (2008)). Similarly, Smith et al. (2019) demonstrate that people use different words to evaluate similar male and female army officers, with males being described as analytical or arrogant, and females as compassionate and temperamental.

Social psychology often utilizes various aspects of attribution to explain behavior, and the measures employed to capture these aspects align accordingly. For instance, the Attribution

Style Questionnaire (Peterson et al. (1982); Dykema, Bergbower, Doctora, and Peterson (1996)) uses hypothetical scenarios to create a self-report instrument that scores individuals' explanatory styles for different events based on three factors: internal versus external, stable versus unstable, and global versus specific causes. The Attributional Complexity Scale is another measure that assesses different attributional constructs (Fletcher, Danilovics, Fernandez, Peterson, and Reeder (1986)). It explores individuals' motivation to understand behavior, their awareness of how interactions with others influence behavior, and their ability to infer internal and external causes of behavior. Similarly, Kelley's co-variation measure (e.g. Kelley (1973)) is a hypothetical self-reported measure that aims to capture aspects such as consensus, consistency, and distinctiveness associated with behavior.

In this article, we take a different approach. Rather than indirectly reflecting and measuring different attributional associated with different theories, we aim to directly examine how males and females are perceived differently and how people perceive different motivations for their actions. To achieve this, we avoid using hypothetical and self-reported measures and instead, address the situation directly and create an incentivized measure. This allows us to explore how individuals judge the actions of males and females more directly and objectively.

To achieve this, we adapted the measure developed by Krupka and Weber (2013) to capture a socially constructed attribution measure. Participants were presented with a specific case in which the participant was lying in the sender and receiver game used in our experiment, which will be further described later. Then, participants were asked to judge the general perception of all participants given the situation, considering five different criteria: "Malicious intention", "Rational calculation", "An emotional decision", "Situational factors", and "An honest mistake". Their task was to guess how all the participants would evaluate this scenario, and if their answer matched the modal response, they received extra compensation. Through this incentivized measure, we aimed to create a social causal attribution measure. The five factors we consider reflect both internal and external causes of behavior, but this is not the focus of our approach. We wanted to explore if males and females are perceived differently based on the various possible explanations given in the scenario.

It's important to note that attribution is often a social construction, involving consensus aspects and creating social inferences based on the context and the individuals involved. Our methodology serves as a proxy for this social construction. By applying it in our experiment, we can directly observe how each gender is perceived differently. This method has several advantages: it allows us to 1) measure attributions directly associated with specific behavior, 2) use an incentivized approach to capture these attributions, and 3) reflect a social construction of perceived attributions. By doing so, we avoid social-desirability bias, as participants are incentivized to think critically about the situation, and the approach measures social perceptions, not directly addressing the participants' bias.

Regarding hypotheses, we do not have a clear expectation for each different measure. Gender stereotypes, previously pointed out in Smith et al. (2019), could suggest that males are more likely to be considered rational, while females are seen as emotional. However, violations of gender roles could also lead to different assessments, as a lying male and a lying female might be considered differently. We summarize this into the following hypothesis:

**H3.3.** There are gender differences in causal attribution regarding engaging in selfish lies.

### 3 Experimental Design

All described hypotheses, the experimental design, and the developed regressions were pre-registered.<sup>1</sup> The study was conducted online using Otree (Chen, Schonger, and Wickens (2016)), and participants were recruited from Prolific. We recruited 240 participants for our study, with 120 participants assigned as senders and the remaining 120 as receivers. Each condition had an equal gender balance, comprising 61 males and 59 females. On average, the sender condition lasted 6 minutes, with participants receiving approximately 2.40 euros, while the receiver condition lasted 13 minutes, with participants receiving approximately 4.50 euros. In the experiment, we used points to represent the money, and 250 points were equal to 1 Pound. The balance table demonstrating that there are no major differences between male and female receivers can be found in Appendix 8.

At the beginning of the experiment, participants provided demographic information, including gender, age, and country. To ensure consistency among participants, we pre-screened them based on specific criteria. All participants met the requirements of residing in the United States, the United Kingdom, or the Netherlands, being aged between 20 and 40, and contributing to a balanced gender distribution in our sample. The pre-screening and matching processes were conducted to ensure a precise alignment of cases and individuals.

We used an adapted version of the sender-receiver game (Capraro (2018); Sánchez-Pagés and Vorsatz (2007b)), with half of the participants playing the role of the sender, and the other half playing the role of the receiver. The sender randomly receives a color, with an 80% chance of receiving green and a 20% chance of receiving blue. The color determined the initial outcomes for both participants. If the color was blue, the sender received 800 points, while the receiver received 200 points. If the color was green, the sender received 400 points, while the receiver received 600 points. The sender's task is to inform the receiver about the color. However, the sender can either truthfully reveal the color or lie by claiming the other color. This means that the points the sender and receiver earn are based on the color the sender communicates, not the color randomly assigned. For example, if the sender was randomly assigned the blue color but claimed to have the green color, the outcome would be based on the green color. The specific instructions for the modified version of the game can be checked in Appendix 6.

The receiver receives the information without being aware of whether it is a lie or not, and observes the demographic information of the sender, including age, gender, and nationality. The sender has to make decisions for 15 different senders. One of the decisions is randomly selected to actually be paid. These cases encompass a total of 10 possible combinations of deceitful behavior, corresponding to all potential combinations of gender, age, and nationality lying (except for Dutch, which had only one age range). Additionally, 5 other cases were included to maintain the correct proportions of lies and truths based on insights from a small pilot study. The pilot study indicated that approximately 20% of individuals received the "blue ball" while around 57% of individuals lied after receiving the "green ball". This led to the inclusion of 15 cases in order to maintain proportions consistent with the 10 cases involving lies.

After completing the 15 decisions, we asked the selected receivers to answer questions about their empirical and normative expectations, as well as questions related to our new measure aimed at investigating gender perception differences. Half of the receivers were asked to provide their opinions on a female participant aged between 20-30 living in the United States, while the other half were asked the same about a male participant with the

---

<sup>1</sup><https://osf.io/qxhg9>

same demographics. To prevent direct comparisons that would overly highlight the gender aspect of the experiment, participants evaluate only one candidate. We randomly selected one measure from the questions and awarded 250 additional points to participants who provided the correct answer. The details regarding this payment are the same as for the elicitation of empirical expectations which is explained in the next paragraph.

Participants were asked to provide their empirical expectations by estimating the frequency of selfish and prosocial lies out of 100 individuals who matched the demographic description of the sender. Specifically, they were asked to estimate the number of individuals claiming that the assigned ball was blue/green when in reality, it was green/blue. The accuracy of their estimate was then compared to the actual frequency observed in the selected group. For every point of distance between the guess and the actual answer, 20 points were deducted, up to 0, from the participant's 250 points.

We utilized the methodology proposed by (Krupka & Weber, 2013) to assess normative expectations. Specifically, we described the scenario in which a participant engages in a selfish lie and requested the participants to evaluate its appropriateness based on a coordination game. Participants were asked to rate various behaviors on a scale from 1 (very socially inappropriate) to 4 (very socially appropriate) and to predict how other participants would evaluate the behavior. An additional 250 points were awarded if the participant correctly selected the most frequently chosen decision among all participants. This measure serves as a proxy for normative expectations by soliciting second-order beliefs regarding the appropriateness levels of specific behavior, thereby creating a social construct while also providing a clear and direct method to incentivize the desired behavior.

To capture and measure these social attributions associated with different genders, we developed a novel methodology building on the approach proposed by (Krupka & Weber, 2013). In this methodology, participants were asked to rate the extent to which they believed others perceived a described selfish lie as being due to five different factors: "Malicious intention", "Rational calculation", "An emotional decision", "Situational factors", and "An honest mistake". Participants provided ratings on a scale from 1 (Strongly disagree) to 4 (Strongly agree) and predicted how others would rate the lying behavior. An additional 250 points were awarded if the participant correctly selected the most frequently chosen decision among all participants for one randomized motivation. This approach allowed us to measure second-order beliefs about the perceived social attribution for lying behavior while reducing social desirability bias and encouraging participation.

## 4 Results

We first present the results associated with lying behavior and then describe the results for punishing possible lies. Additionally, we analyze social norms, empirical and normative expectations, and the new measure for social attributions associated with each gender.

### 4.1 Lies

We start by analyzing the gender differences in lying behavior by using logit regressions. The first regression analyzes the gender differences for selfish lies, the second regression analyzes the prosocial lies. The Male coefficient describes the dummy capturing the gender difference.



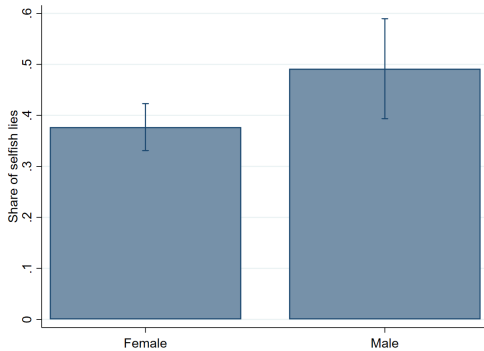


Figure 1: Ratio of selfish lies per gender.

	(1)	(2)
	Selfish Lie	Prosocial Lie
Male	0.468** (0.213)	0.651 (0.517)
Constant	-0.502*** (0.100)	-2.657*** (0.132)
<i>N</i>	120	120

Standard errors clustered by the nationality in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 1: Gender difference on the ratio of selfish/prosocial lies

The results in Table 2 show that males are significantly more likely to engage in selfish lies, with a frequency of 49% compared to 37% for females. While males were also found to be more likely to engage in prosocial lies, the difference was not statistically significant, with a frequency of 12% for males and 7% for females.

**Result 1.1:** No gender differences for prosocial lies

**Result 1.2:** Males performed selfish lies significantly more than females

Therefore, we were able to partially replicate the gender differences observed in previous studies for selfish lies, with males being more likely to lie, but we did not find any significant gender differences for prosocial lies.

## 4.2 Punishment

Our primary analysis focuses on gender differences regarding the likelihood of receiving punishment after engaging in potential selfish lies. In Regression (3), we examine gender differences in receiving punishment for such potential lies while controlling for individual effects. This analysis comprises 10 observations for each of the 120 participants. In Regression (4), we further incorporate nationality and age as control variables to conduct a more comprehensive analysis of gender differences. In Regression (5), we illustrate the gender-based distinctions in punishments associated with potential prosocial lies, again controlling for individual effects. This portion of the analysis is based on 5 observations for each of the 120 participants.

	(3)	(4)	(5)
	Punishment	Punishment	Punishment - Prosocial
Male	0.450*** (0.0535)	0.452*** (0.0535)	0.0420 (0.285)
Age 30-40		-0.258 (0.216)	
U.K.		0.155 (0.130)	
Netherlands		0.285 (0.256)	
Constant	-3.483*** (0.338)	-3.464*** (0.433)	-6.009*** (0.652)
<i>N</i>	1200	1200	600

Standard errors in parentheses  
\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 2: Logit - Punishment

Both regressions (3) and (4) show that males are consistently more punished. On average, females are punished 22% of the time, while males are punished 25%. This difference remains significant even after controlling for age and nationality. Regression (5) shows no significant difference in punishment levels for prosocial outcomes and possible prosocial lies.

**Result 2.1:** No gender differences in receiving punishment for possible prosocial lies

**Result 2.2:** Males are significantly more punished than females for possible selfish lies

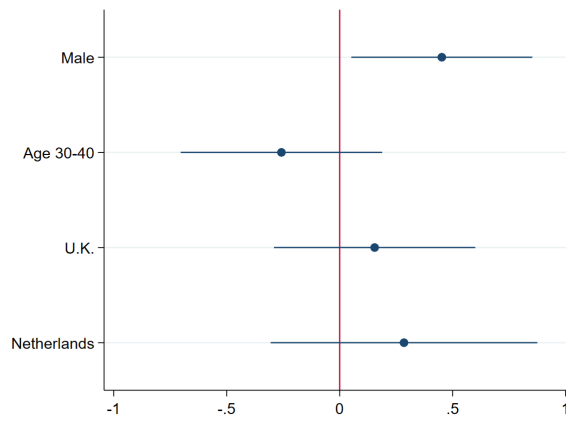


Figure 2: Impact of gender, age, and country in receiving punishment for potential selfish lies

To better understand these differences in punishment, we conducted additional analyses focusing on two key aspects: in-group bias (Male vs. Male/Female) and the potential influence of round sequence on punishment levels. Regression (6) examines the in-group bias by introducing an interaction term between a dummy variable representing the sender's gender and a dummy variable representing the receiver's gender. To identify round trends, we divided our data into two halves and analyzed the interaction between the sender's gender and whether the round was above 7 or not, as observed in regression (7).

	(6)	(7)
	Punishment	Punishment
Male	0.412*** (0.157)	-0.160 (0.116)
Receiver:Male	1.417*** (0.447)	
Male $\times$ Receiver:Male	0.0638 (0.178)	
Second-half		-0.169*** (0.0605)
Male $\times$ Second-half		1.117*** (0.0843)
Constant	-4.147*** (0.593)	-3.453*** (0.364)
$N$	1200	1200

Standard errors clustered by the  
nationality in parentheses  
\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 3: Logit - Share of punishment given in-group bias(6) and trend(7)

Regression (6) does not provide evidence of in-group bias. We observe that when the receiver is male, they tend to punish all participants more frequently than females, as indicated by the coefficient for "Receiver:Male". However, there is no significant difference when they are punishing males versus females, as indicated by the non-significant interaction term. Image 3 illustrate this dynamic:

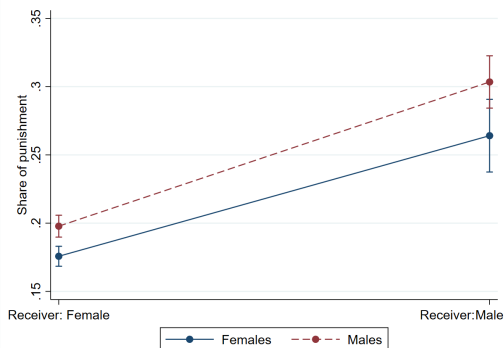


Figure 3: In(out) group bias - sender and receiver genders

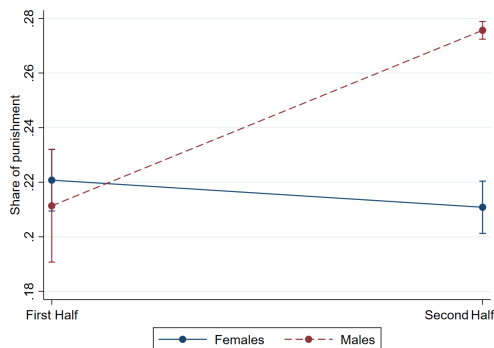


Figure 4: Round dynamics - first and second half

Meanwhile, regression (7) reveals that the number of rounds plays a significant role in gender differences. In the initial part of the experiment, there is no significant gender difference in punishment levels, with females being even slightly more punished (males - 21% and females - 22%). However, in the second half of the experiment, the difference becomes highly significant. Approximately 28% of males are being punished, whereas there is no significant difference observed for females (at around 21%). This tendency might possibly be explained by the dynamics of observing possible lies. Participants seem to trust that the behavior is not a lie during the very first interactions, even if the likelihood of a lie is high, but they start to consider the possibility of lying more seriously when the behavior is observed more frequently. As participants begin to wonder more often about possible lies, they also start to punish males more regularly.

### 4.3 Social Norms and Attributions

Besides the behavioral changes, we also elicited the receivers' gender perceptions and their empirical and normative expectations. First, we analyze if the sender's gender affects any of these measures. Subsequently, we examine if any of these factors affect the decision-making process.

To start, we focus on analyzing empirical expectations. Specifically, we asked the participants to envision 100 individuals fitting a particular sociodemographic description and determine how many of them would engage in a selfish lie or a prosocial lie. Therefore, each of the 120 participants indicated their expectations for one individual. Linear regression (8) examines the gender differences related to selfish lies, and linear regression (9) assesses the gender differences regarding prosocial lies.

	(8)	(9)
	Selfish Lies	Prosocial Lies
Male	8.155*	-2.250
	(4.432)	(4.392)
Constant	49.08***	21.64***
	(3.108)	(3.079)
<i>N</i>	120	120

Standard errors in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 4: Linear regression: differences in empirical expectations by gender

Regression (8) reveals that males are expected to engage in more selfish lies, exhibiting an 8% point difference. Specifically, females are expected to lie approximately 49% of the time, while males are anticipated to lie around 57% of the time. On the other hand, regression (9) indicates that males are expected to perform fewer prosocial lies, with 19% of individuals expecting males to engage in such behavior, compared to 22% for females.

**Result 3.1a:** Males are expected to perform more selfish lies than females. **Result 3.1b:** Males are expected to perform fewer prosocial lies than females.

We shift our focus to the normative expectations, which were assessed using Krupka and Weber (2013)'s measure - ranging from 1 (Very Socially Inappropriate) to 4 (Very Socially Appropriate). In order to examine gender differences in normative expectations, we conducted an ordered logit regression. Regression (10) captures explicitly the gender differences for a selfish lie, while regression (11) captures the gender differences for truthfully reporting the selfish outcome.

	(10)	(11)
	Selfish Lie	Selfish Truth
Male	0.158	0.232
	(0.344)	(0.400)
<i>N</i>	120	120

Standard errors in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 5: Ordered logit: differences in normative expectations by gender

Lies and truths also manifested in terms of appropriateness levels. Selfish lies are reported to have appropriateness levels around 1.9, whereas truthfully revealing the selfish outcome has appropriateness levels around 3.6. However, no significant difference is observed across genders, as males and females are judged similarly for both selfish lies and truthfully reporting the selfish outcome.

**Result 3.2a:** No gender differences when performing a selfish lie than females.

**Result 3.2b:** No gender differences when truthfully revealing a selfish outcome.

Participants were also asked to rate selfish lies based on the possible motivations behind them. They used a scale ranging from 1 (strongly disagree) to 4 (strongly agree) to indicate whether they believed other participants would consider the motive for a selfish lie as a "malicious intention" (Regression 12), "rational calculation" (Regression 13), "emotional decision" (Regression 14), "situational factor" (Regression 15), or "honest mistake" (Regression 16). The following results reflect the findings from ordered logit regressions in Table 7, the average score by gender is observed in Table 6:

	Males	Females
Malicious intention	2.85 (0.86)	2.87 (0.89)
Rational calculation	3.14 (0.87)	2.97 (0.95)
Emotional decision	2.37 (0.93)	2.33 (0.95)
Situational factor	2.92 (0.86)	2.87 (0.90)
Honest mistake	1.55 (0.79)	1.46 (0.77)
Observations	59	61

Standard deviation in parentheses

Table 6: Gender-wise Average Attribution Score for Lies

	(12) Malicious	(13) Rational	(14) Emotional	(15) Situational	(16) Mistake
Male	-0.0291 (0.336)	-0.587* (0.345)	-0.332 (0.333)	-0.513 (0.346)	0.348 (0.365)
<i>N</i>	120	120	120	120	120

Standard errors in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 7: Ordered logit: differences in social attributions by gender

**Result 3.3:** We observed a significant difference in the attribution levels of rational calculation, with females being attributed higher levels of rational calculation.

The results indicate that gender might contribute to the formation of perceptions about various motivations behind selfish lies. A statistically significant result is associated with the "rational calculation" motive, with females being attributed higher levels in that regard, while the distinctions for other motives are relatively subtle and not statistically significant.

## 5 Discussion

Our study aims to shed light on punishment gaps, which describe how minorities and underprivileged groups are disproportionately punished and judged for mistakes or misconduct. Various studies have observed these patterns in real-life data (e.g., Egan et al. (2022); Gupta, Mortal, Silveri, Sun, and Turban (2020)), but these observations cannot fully disentangle the actual causes. Other studies employ hypothetical and specific scenarios (e.g., Berinsky, Hutchings, Mendelberg, Shaker, and Valentino (2011); Sommers and Ellsworth (2000)), where individuals might respond differently than when facing real consequences. In our experiment, we create a controlled setting with real monetary consequences to investigate whether discriminatory punishment towards underprivileged groups exists. Furthermore, we seek to understand potential causes for such behavior.

We designed a sender-receiver game enabling senders to lie about the outcomes, breaching ethical norms. Receivers observe senders' actions without certainty about their veracity, while also receiving information about the sender's demographics, including gender. Subsequently, senders decide whether to penalize potential liars. The setting creates a substantial likelihood of deception (around 66% of selfish outcomes are lies), and our objective is to check if people exhibit a higher tendency to penalize males or females for potential lies. For the purpose of our experiment, females represent an underprivileged group, and a greater level of punishment would indicate a punishment gap. We then delve into empirical and normative expectations, along with gender-based causal attributions, to clarify any behavioral differences.

Firstly, our inquiry revolves around the propensity of each gender to engage in selfish lies, with an attempt to replicate findings from various sources (e.g., Dreber and Johannesson (2008); Friesen and Gangadharan (2012); Erat and Gneezy (2012); Grosch and Rau (2017); Capraro and Peltola (2018); Cappelen et al. (2013); Biziou-van Pol et al. (2015)). Our results suggest that males exhibit a higher likelihood of engaging in selfish lies, accounting for 49% of cases, compared to females at 37%. It is important to note that results may not consistently hold strong, as various regression setups, including cluster adjustments, could lead to statistically insignificant outcomes. Additionally, we note a trend where males are more prone to prosocial lies (12% males vs. 7% females), but this difference lacks significance.

Our study's primary focus was to analyze gender-based disparities in punishments for potential selfish lies. Overall, our observations revealed a subtle yet statistically significant difference: males received punishment slightly more often —25% of the time— compared to females - 23% of the time. We also try to understand behavioral nuances associated with this difference. We investigate whether in-group bias as well as the shared gender of the sender and receiver impacts the dynamics, and influences the punishment levels. Additionally, we explore the role of the number of rounds in shaping punitive actions.

Our analysis does not show evidence of in-group or out-group bias, and both genders tend to act similarly when confronted with interactions involving the same or different genders. However, we observe that males punish more often than females; males consistently punish around 10 percentage points more often than females (28.4% for males vs. 18.6% for females). Future research might aim to understand whether there are additional sources and contexts through which in-group behavior could impact levels of punishment.

Meanwhile, the rounds of the experiment exert an influence on punishment patterns. Receivers have to make decisions for 15 different senders in a row. In the experiment's initial rounds, participants displayed a balanced inclination to punish males and females (21% for

males vs. 22% for females), with females being slightly more likely to be punished, although not statistically significant. However, a shift occurred in the second half of the experiment. Receivers began punishing males more frequently, with males facing punishment in 28% of instances, while females experienced punishment in 21% of instances.

This emerging trend possibly reflects the importance of uncertainty alongside a decline in trust. Initially, participants trusted the others, resulting in fewer penalties. As the experiment progressed, the prevalence of selfish outcomes led to caution against deception, driving up penalties. A participant's post-experiment remark exemplified this: "I assumed some of the blue ball picks were legitimate, but began to feel that there were too many for such a low chance."

In summary, our results indicate that the underprivileged group – females – faced fewer penalties than males, hence we observed no direct evidence of significant punishment gaps. Our experimental setup was carefully designed to incorporate uncertainty, mirroring real-life situations where individuals often lack clarity about the actual circumstances. The findings suggest two influences: 1) uncertainty and 2) a higher likelihood of male deception. The existence of discrimination, as seen in the punishment gap, can be attributed to various factors. Becker (2010) describes a model in which discrimination can stem from statistical reasoning (involving beliefs about differing behavior among groups) or taste-based preferences (favoring one group over another). Our uncertain setting may impact outcomes. Initially, both genders were treated similarly. However, the prevalence of selfish outcomes eroded this trust, fostering skepticism. As a result, individuals leaned toward penalizing those perceived as deceivers, aligning with the higher likelihood of male deception. Subsequently, our analysis will assess whether these beliefs translate into actual observations.

We analyzed participants' empirical expectations upon the experiment's conclusion, which reveal real behaviors and punitive outcomes. Senders consistently described a higher likelihood of lying to males (57%) compared to females (49%). This alignment supports the idea that uncertainty and eroded trust lead individuals to speculate about deception-prone individuals, subsequently influencing their punitive decisions towards males.

We also examined participants' normative expectations to understand their judgments of such actions. Participants do associate the lying behavior with lower levels of appropriateness, with lies scoring around 1.9, while truthfully reporting the selfish outcome is associated with a higher score of 3.2. No major gender differences were observed. Females scored lower levels in both situations, being perceived as less appropriate when lying or truthfully reporting the selfish outcome; however, this difference is not statistically significant. The lack of significance might be attributed to the smaller sample size associated with the survey at the end of the experiment, resulting in around 60 observations for each gender. Future research could further investigate this possibility.

Similarly, we delved into social causal attributions to discern differences in perceived motivations behind lies and checked for distinctions in how people perceive males and females differently. Our newly developed method captures attributions on a scale of 1 to 4, where higher scores reflect greater degrees of attribution. The results indicate that people primarily attribute lies to "Rational calculation" (3.14) and "Situational factors" (2.92), with the lowest attribution given to "Honest mistake" (1.55). The results also reveal disparities in attributions between males and females. Senders were more inclined to assign higher scores to "Rational calculations" for females rather than males. Other minor differences were observed, but they were not statistically significant. For instance, males scored higher levels in "Honest mistake". These outcomes potentially illustrate stronger scrutiny of females when deviating from their gender norms. In that regard, a lying male might be perceived



as typical, given the perceived likelihood to lie, whereas a lying female could be seen as a violator, engaging in intentional deceptive reasoning, which reflects a direct violation of gender expectations, e.g. Diekmann and Eagly (2008).

Hence, our new method has the potential to capture how people perceive various situations and can be adapted to different settings. Furthermore, it enhances the exploration of social norms. The method can also be easily adapted to include a direct and incentivized approach to capture motivated reasoning (e.g., Epley and Gilovich (2016)) or social image concerns (e.g., Foerster and van der Weele (2021)). Future research is necessary to explore the method's opportunities and limits and to further understand how individuals perceive the situations, directly relating it with behavior.

In conclusion, within this lying context, we do not observe a punishment gap, and males are more likely to receive punishment. The results also align with empirical expectations and actual lying behavior, as males are both more likely to lie and expected to do so more often. However, these results reflect the specific context where uncertainty exists regarding the likelihood of unethical violation.

In the initial rounds, males and females are treated similarly, despite the higher likelihood of males lying. Additionally, there are subtle differences in how people perceive lying behavior based on gender. For instance, females are perceived as more "rational" for exhibiting the same behavior. It's important to note that different contexts and situations may lead to the emergence of a punishment gap. For example, when the ethical violation is known, these minor differences might have a greater impact, resulting in a higher punishment for females. Future research could further analyze the relationship within varying contexts.

## 6 Conclusion

Our study delved into the phenomenon of punishment gaps, wherein minority and or underprivileged groups face disproportionate penalties for mistakes or misconduct, in a controlled and incentivized setting. We found that males were more likely to engage in selfish lies and were more likely to get punished for the potential lies. Uncertainty and declining trust, as the prevalence of selfish outcomes becomes salient, appear to drive this difference in punishment levels. These results reflect the participants' empirical expectations, as participants indicate that they expect males to be more likely to lie.

Additionally, we have developed a new measure to capture attributions associated with actions. This new measure reveals that females are attributed with higher levels of 'rational calculation' towards their lies, highlighting differences in how people perceive males and females for the same action.

The punishment gap is not universally present. The context and the specific behavior playing a role are important factors that need to be further studied to better understand this situation.

## References

- Becker, G. S. (2010). *The economics of discrimination*. University of Chicago press.
- Berinsky, A. J., Hutchings, V. L., Mendelberg, T., Shaker, L., & Valentino, N. A. (2011). Sex and race: Are black candidates more likely to be disadvantaged by sex scandals? *Political Behavior*, 33, 179–202.

- Beyer, S. (1998). Gender differences in causal attributions by college students of performance on course examinations. *Current psychology*, *17*, 346–358.
- Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.
- Biziou-van Pol, L., Haenen, J., Novaro, A., Liberman, A. O., & Capraro, V. (2015). Does telling white lies signal pro-social preferences? *Judgment and Decision Making*, *10*(6), 538–548.
- Brañas-Garza, P., Capraro, V., & Rascon-Ramirez, E. (2018). Gender differences in altruism on mechanical turk: Expectations and actual behaviour. *Economics Letters*, *170*, 19–23.
- Brandts, J., & Charness, G. (2003). Truth or consequences: An experiment. *Management Science*, *49*(1), 116–130.
- Brescoll, V. L., & Uhlmann, E. L. (2008). Can an angry woman get ahead? status conferral, gender, and expression of emotion in the workplace. *Psychological science*, *19*(3), 268–275.
- Burnham, T. C. (2018). Gender, punishment, and cooperation: Men hurt others to advance their interests. *Socius*, *4*, 2378023117742245.
- Cappelen, A. W., Konow, J., Sørensen, E. Ø., & Tungodden, B. (2013). Just luck: An experimental study of risk-taking and fairness. *American Economic Review*, *103*(4), 1398–1413.
- Capraro, V. (2018). Gender differences in lying in sender-receiver games: A meta-analysis. *Judgment and Decision making*, *13*(4), 345–355.
- Capraro, V., & Peltola, N. (2018). Low cognitive reflection predicts honesty for men but not for women. *arXiv preprint arXiv:1805.08316*.
- Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.
- Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.
- Crosan, R., & Buchan, N. (1999). Gender and culture: International experimental evidence from trust games. *American Economic Review*, *89*(2), 386–391.
- Diekmann, A. B., & Eagly, A. H. (2008). Of men, women, and motivation. *Handbook of motivation science*, *434*.
- Dimant, E., & Gesche, T. (2021). Nudging enforcers: How norm perceptions and motives for lying shape sanctions.
- Dittrich, M. (2015). Gender differences in trust and reciprocity: evidence from a large-scale experiment with heterogeneous subjects. *Applied Economics*, *47*(36), 3825–3838.
- Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters*, *99*(1), 197–199.
- Dykema, J., Bergbower, K., Doctora, J. D., & Peterson, C. (1996). An attributional style questionnaire for general use. *Journal of Psychoeducational Assessment*, *14*(2), 100–108.
- Eckel, C. C., & Grossman, P. J. (1996). The relative price of fairness: Gender differences in a punishment game. *Journal of Economic Behavior & Organization*, *30*(2), 143–158.
- Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics*, *9*(2), 1–22.

- Egan, M., Matvos, G., & Seru, A. (2022). When harry fired sally: The double standard in punishing misconduct. *Journal of Political Economy*, *130*(5), 1184–1248.
- Eisenkopf, G., Gurtoviy, R., & Utikal, V. (2017). Punishment motives for small and big lies. *Journal of Economics & Management Strategy*, *26*(2), 484–498.
- Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic perspectives*, *30*(3), 133–140.
- Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723–733.
- Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, *362*(6412), eaas9899.
- Fišar, M., Kubák, M., Špalek, J., & Tremewan, J. (2016). Gender differences in beliefs and actions in a framed corruption experiment. *Journal of Behavioral and Experimental Economics*, *63*, 69–82.
- Fletcher, G. J., Danilovics, P., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity scale. *Journal of Personality and Social Psychology*.
- Foerster, M., & van der Weele, J. J. (2021). Casting doubt: Image concerns and the communication of social impact. *The Economic Journal*, *131*(639), 2887–2919.
- Friesen, L., & Gangadharan, L. (2012). Individual level evidence of dishonesty and the gender effect. *Economics Letters*, *117*(3), 624–626.
- García-Gallego, A., Georgantzís, N., & Jaramillo-Gutiérrez, A. (2012). Gender differences in ultimatum games: Despite rather than due to risk attitudes. *Journal of Economic Behavior & Organization*, *83*(1), 42–49.
- Garcia-Retamero, R., & Lopez-Zafra, E. (2009). Causal attributions about feminine and leadership roles: A cross-cultural comparison. *Journal of Cross-Cultural Psychology*, *40*(3), 492–509.
- Grosch, K., & Rau, H. A. (2017). Gender differences in honesty: The role of social value orientation. *Journal of Economic Psychology*, *62*, 258–267.
- Gupta, V. K., Mortal, S. C., Silveri, S., Sun, M., & Turban, D. B. (2020). You're fired! gender disparities in ceo dismissal. *Journal of Management*, *46*(4), 560–582.
- Heider, F. (2013). *The psychology of interpersonal relations*. Psychology Press.
- Heilman, M. E., & Chen, J. J. (2005). Same behavior, different consequences: reactions to men's and women's altruistic citizenship behavior. *Journal of Applied Psychology*, *90*(3), 431.
- Jung, S., & Vranceanu, R. (2015). Gender interaction in teams: Experimental evidence on performance and punishment behavior. *Available at SSRN 2626327*.
- Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107.
- Kouchaki, M., & Smith, I. H. (2014). The morning morality effect: The influence of time of day on unethical behavior. *Psychological science*, *25*(1), 95–102.
- Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.
- Laske, K., Saccardo, S., & Gneezy, U. (2018). Do fines deter unethical behavior? the effect of systematically varying the size and probability of punishment. *The effect of systematically varying the size and probability of punishment (April 5, 2018)*.
- Mieth, L., Buchner, A., & Bell, R. (2017). Effects of gender on costly punishment. *Journal of Behavioral Decision Making*, *30*(4), 899–912.
- Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature*, *56*(3), 799–866.

- Peeters, R., Vorsatz, M., & Walzl, M. (2013). Truth, trust, and sanctions: on institutional selection in sender–receiver games. *The Scandinavian Journal of Economics*, *115*(2), 508–548.
- Peterson, C., Semmel, A., Von Baeyer, C., Abramson, L. Y., Metalsky, G. I., & Seligman, M. E. (1982). The attributional style questionnaire. *Cognitive therapy and research*, *6*(3), 287–299.
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: examining reasons for acting in two domains. *Journal of personality and social psychology*, *57*(5), 749.
- Saad, G., & Gill, T. (2001). Sex differences in the ultimatum game: An evolutionary psychology perspective. *Journal of Bioeconomics*, *3*, 171–193.
- Sánchez-Pagés, S., & Vorsatz, M. (2007a). An experimental study of truth-telling in a sender–receiver game. *Games and Economic Behavior*, *61*(1), 86–112.
- Sánchez-Pagés, S., & Vorsatz, M. (2007b). An experimental study of truth-telling in a sender–receiver game. *Games and Economic Behavior*, *61*(1), 86–112.
- Sánchez-Pagés, S., & Vorsatz, M. (2009). Enjoy the silence: an experiment on truth-telling. *Experimental Economics*, *12*(2), 220–241.
- Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals. *Job Market Paper*, 141–145.
- Shaver, K. G. (2016). *An introduction to attribution processes*. Routledge.
- Smith, D. G., Rosenstein, J. E., Nikolov, M. C., & Chaney, D. A. (2019). The power of language: Gender, status, and agency in performance evaluations. *Sex Roles*, *80*, 159–171.
- Sommers, S. R., & Ellsworth, P. C. (2000). Race in the courtroom: Perceptions of guilt and dispositional attributions. *Personality and Social Psychology Bulletin*, *26*(11), 1367–1379.
- Visser, M. S., & Roelofs, M. R. (2011). Heterogeneous preferences for altruism: Gender and personality, social status, giving and taking. *Experimental Economics*, *14*(4), 490.

## Appendix

### Balance Table

	Males	Females	Difference
	Mean	Mean	Difference
Age	31.02 (5.49)	29.32 (6.32)	-1.69 [0.12]
Residence	2.49 (1.50)	2.81 (1.92)	0.32 [0.31]
Brother	1.13 (0.34)	1.12 (0.33)	-0.01 [0.84]
Observations	61	59	120

\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

Standard deviation in parentheses

t statistics in brackets

Table 8: Balance Table

### Difference given nationality

	U.S.	U.K.	Netherlands
	Mean	Mean	Mean
Selfish Lie	0.45 (0.50)	0.50 (0.58)	0.14 (0.38)
Prosocial Lie	0.08 (0.28)	0.00 (0.00)	0.29 (0.49)
Observations	109	4	7

Standard deviation in parentheses

Table 9: Differences on behavior given residence

The majority of our senders are from the U.S., and there are not many people from the U.K. and the Netherlands. People from the Netherlands acted quite differently from the others. We decided to cluster the behavior on a nationality level, and the result of lying is not always robust when considering other specifications. However, the other results are consistent, and all the results show no clusters present next.

## Robustness check

	(1)	(2)
	Selfish Lie	Prosocial Lie
Male	0.468 (0.371)	0.651 (0.655)
Constant	-0.502* (0.264)	-2.657*** (0.517)
<i>N</i>	120	120

Standard errors in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 10: Gender differences on lying with no clustering

The difference is not significant without clustering.

	(3)	(4)
	Punishment	Punishment
Male	0.450** (0.027)	0.452** (0.027)
30-40		-0.258 (0.257)
UK		0.155 (0.496)
Netherlands		0.285 (0.345)
Constant	-3.483*** (0.000)	-3.464*** (0.000)
<i>N</i>	1200	1200

*p*-values in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Table 11: Gender differences on punishment with no clustering

The difference is robust to the changes in the clustering.

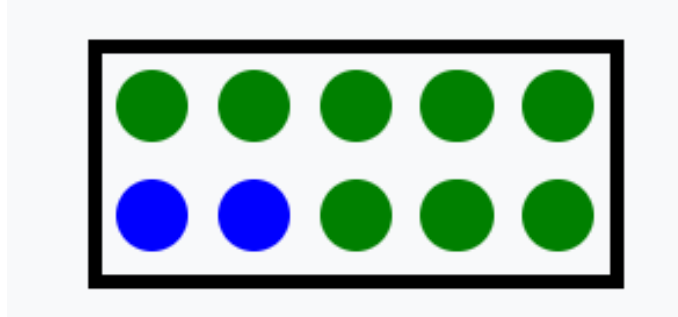
## Instructions

### Part 1 - Participant 1

In this study, you will be randomly paired with another participant. Either you or the other participant will be randomly selected as 'Participant 1' or 'Participant 2'.

**You were selected as Participant 1/2.**

The task is divided into two parts.  
Consider the box below:



- 8 balls are **GREEN** (80% chance).
- 2 balls are **BLUE** (20% chance).

A ball will be randomly picked by the computer and shown to Participant 1. The randomized ball's color establishes the number of points that each participant **starts** with.

- A **GREEN** ball means that Participant 1 starts with **400 points**, while Participant 2 starts with **600 points**.
- A **BLUE** ball means that Participant 1 starts with **800 points**, while Participant 2 starts with **200 points**.

After Participant 1 has seen the color of the ball, Participant 1 has to inform Participant 2 of the color of the ball. The actual color informed by Participant 1 determines the final number of points that each participant receives.

Notice that Participant 1 can correctly inform Participant 2 about the observed color or may decide to misinform and untruthfully report the ball's color. That is:

- If Participant 1 informs Participant 2 that the color was **GREEN**, Participant 1 gets **400 points** and Participant 2 gets **600 points**, even if the ball is **BLUE**.
- If Participant 1 informs Participant 2 that the color was **BLUE**, Participant 1 gets **800 points** and Participant 2 gets **200 points**, even if the ball is **GREEN**.

## Part 2 - Participant 2

After Participant 1 has communicated the ball color, it is Participant 2's turn.

Participant 2 receives the demographic information that Participant 1 answered on the previous screen.

Participant 2 receives the colors of the balls informed by Participant 1 during part 1.

Notice that Participant 2 does not know the actual color that was randomly picked by the computer and shown to Participant 1. Participant 2 only knows the information sent by Participant 1.

After being informed about the color of the ball, Participant 2 has the option to punish Participant 1. Punishment is not mandatory.

If Participant 2 decides not to punish, the participant's payoffs stay the same.

If Participant 2 decides to punish Participant 1, Participant 1 loses 300 points and Participant 2 loses 50 points.

### Earnings

In the experiment, you can earn points. The points you earn will be converted into Pounds at the rate:

$$250 \text{ points} = 1.00 \text{ Pounds.}$$

Your extra earnings will be added to your account a couple of days after the end of the experiment.

### Experiment:

#### Information

**You were selected as Participant 2**

In the following pages, you will see the information sent by 15 participants who took the role of Participant 1.

All decisions are real, made by other participants. One of them is the one you were randomly paired with. We will use the decision associated with this participant to calculate your and Participant 1's payoffs.

You will have to decide whether to punish Participant 1 or not to punish depending on the color of the ball that Participant 1 has informed you about.

Remember, you do not know which color was drawn by the computer and shown to Participant 1. You only know what Participant 1 has informed you. Participant 1 may or may not have untruthfully reported the color that was drawn by the computer.

Figure 5: Extra information for the Receiver




### Make your choice - 1

This **Participant 1** has the following characteristics:

**Lives in:** United Kingdom  
**Gender:** Female  
**Age between:** 20-30

This Participant 1 informed that the ball was:

  
**A BLUE ball.**

Participant 1 is getting **800 points**. You are getting **200 points**.

What will you do?

Do nothing:  
Participant 1 gets **800 points**. You get **200 points**.

Punish Participant 1:  
Participant 1 gets **500 points**. You get **150 points**.

[Next](#)


Figure 6: Example of decision screen

### Make your choice - 1

This **Participant 1** has the following characteristics:

**Lives in:** United Kingdom  
**Gender:** Female  
**Age between:** 20-30

This Participant 1 informed that the ball was:

  
**A BLUE ball.**

Participant 1 is getting **800 points**. You are getting **200 points**.

What will you do?

Do nothing:  
Participant 1 gets **800 points**. You get **200 points**.

Punish Participant 1:  
Participant 1 gets **500 points**. You get **150 points**.

[Next](#)

Figure 7: Example of decision screen

## Survey Instructions

The following survey is divided into three parts.

In part 1, we will ask you to state your beliefs about the behavior of Participants 1.

In part 2, we will ask you to evaluate how the other participants in this experiment assess the decision situation of Participant 1.

In part 3, we will ask you for your own perceptions of the decision situation.

We will randomly select one question from parts 1 and 2. For this question, you can earn a **bonus of 250 points**. The closer your answer is to the correct one, the higher your earnings.

Further information about how the earnings are determined will be given to you on the following pages.

[Next](#)

Figure 8: Start of belief elicitation

### Part 1 - What did the others do?

In the following tasks, you are asked to describe your beliefs about the participants 1 on the previous task. We will describe the features of the participants you will evaluate. Pay attention on that information.

To determine the correct answer, we will measure the average decisions of all participants 1 that fit the reported description.

The closer your answer is to the average decision of all participants 1 the higher your payoff.

A correct answer is rewarded with 250 points. 20 points are subtracted from these 250 points for each unit your answer departs from the average decision.

That is, if the average behavior is "X", and your guess is "X+2"; you will earn 210 extra points if this question was the one randomly selected.

[Next](#)

Figure 9: Instructions empirical expectations

## Survey 1

Consider 100 participants acting as Participant 1 in the previous task. All of them have the following characteristics:

This **Participant 1** has the following characteristics:

**Lives in:** U.S.  
**Gender:** Female  
**Age between:** 20-30

If you want to look up any detail about the task, you can check the instruction below.

Remember:

- If Participant 1 informed that the ball was **GREEN**: Participant 1 **400 points**, Participant 2 **600 points**
- If Participant 1 informed that the ball was **BLUE**: Participant 1 **800 points**, Participant 2 **200 points**

For this part, you have to evaluate the behaviors of Participant 1.

**Remember:** A correct answer is rewarded with 250 points. 20 points are subtracted from these 250 points for each unit your guess departs from the actual answer.

**Question 1:** From those 100 individuals, how many would falsely inform Participant 2 that the color was **BLUE** when the selected color was **GREEN**?

**Question 2:** From those 100 individuals, how many would falsely inform Participant 2 that the color was **GREEN** when the selected color was **BLUE**?

Next

Figure 10: Empirical expectations

## Part 2 - What is your opinion on what the others perceive?

In the following tasks, you will be asked to describe your opinion on the other participants' perceptions.

Your goal is to guess the **most common response provided by the participants taking part in this experiment**.

For example:

You will be asked: "How appropriate do others think that the following behavior is on a scale from 1 to 4"

We will measure the **most common response of all participants on each question to determine the correct answer**.

You will earn 100 points if your response to a randomly-selected question is the same as the most common response.

Remember, we will randomly select one question from parts 1 and 2 according to which you will be paid out.

Next

Figure 11: Instructions normative expectations

## Survey 2

Consider other participants acting as Participant 1 in the previous task. All of them have the following characteristics.

This **Participant 1** has the following characteristics:

**Lives in:** U.S.  
**Gender:** Female  
**Age between:** 20-30

Remember:

- If Participant 1 gave the information that the ball was **GREEN**: Participant 1 **400 points**, Participant 2 **600 points**
- If Participant 1 gave the information that the ball was **BLUE**: Participant 1 **800 points**, Participant 2 **200 points**

For each behavior, you have to evaluate how the other participants of the experiment evaluate the situation. That is, you have to guess what the **other participants** think about the situation from "very socially inappropriate" (1) to "very socially appropriate" (4).

The first column represents what Participant 1 randomly received. The second column indicates which information Participant 1 gave Participant 2.

**Remember:** You will earn 100 points if your response to a randomly-selected question is the same as the most common response.

**Question 1: How appropriate is the described situation according to the other participants?**

P1 Received		P1 informed	
GREEN		BLUE	
"Very Socially Inappropriate"	"Somewhat Socially Inappropriate"	"Somewhat Socially Appropriate"	"Very Socially Appropriate"
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Question 2: How appropriate is the described situation according to the other participants?**

P1 Received		P1 informed	
BLUE		BLUE	
"Very Socially Inappropriate"	"Somewhat Socially Inappropriate"	"Somewhat Socially Appropriate"	"Very Socially Appropriate"
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Next](#)

Figure 12: Normative expectations

### Survey 3

Consider other participants acting as Participant 1 in the previous task. All of them have the following characteristics.

This **Participant 1** has the following characteristics:

**Lives in:** U.S.  
**Gender:** Female  
**Age between:** 20-30

Remember:

- **GREEN** : Participant 1 **400 points**, Participant 2 **600 points**
- **BLUE** : Participant 1 **800 points**, Participant 2 **200 points**

For each behavior, please evaluate how the others perceive the motivation behind the behavior.

Describe from **"Strongly Disagree" (1)** to **"Strongly agree" (4)** your beliefs on how the others perceive the behavior according to the description on the third column.

**The first column represents the color Participant 1 randomly received. The second column indicates which information Participant 1 gave Participant 2. The third column represents possible motives behind the behavior of Participant 1.**

For example: if the third column describes "situational factor", you have to answer "to what extent the other participants perceive the behavior described is a due to a -situational factor-".

**Remember:** You will earn 100 points if your response to a randomly-selected question is the same as the most common response.

Figure 13: Instructions attribution

Question 1: To what extent the other participants perceive the behavior described is a due to the motive bellow?

P1 Received	P1 informed	Motive	
GREEN	BLUE	Malicious intention	
"Strongly disagree"	"Partially disagree"	"Partially agree"	"Strongly agree"
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 14: Example of attribution question