CESifo WORKING PAPERS

8564
2020

September 2020

Large Losses from Little Lies: Randomly Assigned Opportunity to Misrepresent Substantially Lowers Later Cooperation and Worsens Income Inequality

Michalis Drouvelis, Jennifer Gerson, Nattavudh Powdthavee, Yohanes E. Riyanto

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8564*

Large Losses from Little Lies: Randomly Assigned Opportunity to Misrepresent Substantially Lowers Later Cooperation and Worsens Income Inequality

Abstract

Social media has made anonymized behavior online a prevalent part of many people's daily interactions. The implications of this new ability to hide one's identity information remain imperfectly understood. Might it be corrosive to human cooperation? This paper investigates the possibility that a small deceptive act of misrepresenting some information about one's real identity to others – a social media-related behavior commonly known as 'catfishing' – increases the likelihood that the individual will go on to behave uncooperatively in an otherwise anonymous prisoner's dilemma game. In our intention-to-treat analysis, we demonstrated that randomly allowing people to misrepresent their gender identity information reduced the aggregate cooperation level by approximately 12-13 percentage points. Not only that the average catfisher was substantially more likely to go on to defect than participants in the control and the true gender groups, those who were paired with a potential catfisher also defected significantly more often as well. Participants also suffered a significant financial loss from having been randomly matched with a catfisher; 64% of those who played against someone who chose to misrepresent information about their gender received a payoff of zero from the prisoner's dilemma game. Our results suggest that even small short-term opportunities to misrepresent one's identity to others can potentially be extremely harmful to later human cooperation and the economic well-being of the victims.

JEL-Codes: C920, D910.

Keywords: cooperation, misrepresentation, social media, social dilemma, experiment.

Michalis Drouvelis
University of Birmingham / UK
M.Drouvelis@bham.ac.uk

Jennifer Gerson
City, University of London / UK
jennifer.gerson.4@city.ac.uk

*Nattavudh Powdthavee**
Warwick Business School / UK
Nattavudh.powdthavee@wbs.ac.uk

Yohanes E. Riyanto
Nanyang Technological University/ Singapore
yeriyanto@ntu.edu.sg

*corresponding author

Author contributions: M.D., J.G., N.P. and Y.E.R. designed and conducted the experiments; N.P. analyzed data; and M.D., J.G., N.P. and Y.E.R. wrote the paper.

Acknowledgement: We are grateful to Ta Vejpattarasiri and Erwin Wong for their help with the data collection. We are also thankful to Andrew Oswald, Carol Graham, and Daniel Sgroi for their excellent comments on the draft. The project was funded by M.D., N.P., and Y.E.R.'s annual personal research budget from their respective universities. The experiment was approved by the HSSREC ethics board at the University of Warwick (Ref: HSS 75/18-19).

1. Introduction

The invention of social media has forever transformed the way humans are connected. For millions of individuals, social interactions now take place entirely remotely online. Many people frequently use different social media platforms to buy and sell products, engage in romantic relationships, and exchange new ideas with other people from around the world, many of whom are genetically unrelated strangers whose real identity may be completely different from their online persona. This raises the question of whether social media, which improves human connection but also allows people to project whomever they want to be to others, enhanced or hindered human cooperation, which is vital for the functioning of society that heavily relies on social media information systems (SMIS) to operate.

While there is currently little insight into this question, we know from a large body of research from across different disciplines that, unlike other creatures, people frequently cooperate with strangers in large groups, with people whom they will never meet again, and often in situations where reputation gains are small or absent (Fehr & Gächter, 2000, 2002; Hermann, Thöni & Gächter, 2008; Rand & Nowak, 2013). There is also considerable evidence that the ability to send and receive communication, both face-to-face and verbally while maintaining anonymity through a chat room, often improves human's capacity to work together for socially beneficial outcomes (Miller, Butts & Rode, 2002; Bochet, Page & Putterman, 2006; Balliet, 2010). This is mainly because communication allows conditional cooperators – i.e., those who are willing to cooperate if they expect others to cooperate as well – to reveal themselves to each other, which has been shown to improve the average cooperation in many settings in the past (e.g., Fischbacher, Gächter & Fehr, 2001; Frey and Meier, 2004; Chaudhuri & Paichayontvijit,

2006).[1] It might, therefore, be natural to assume that social media, with its ability to connect strangers and allows them to communicate with each other, has further enhanced cooperation even in situations in which doing so is not consistent with their material self-interest.

However, casual evidence as well as daily experience suggest that uncooperative behaviors on social media may be more widespread than previous findings on cooperation would have suggested. Take the more extreme statistics on online frauds, for example. According to the Federal Trade Commission (FTC), there were over 25,000 victims who filed a report about romance scams in 2019, with a total loss of $201 million going to the scammers (FTC, 2020). Another study has estimated social media-enabled cybercrimes to cost the global economy at least $3.25 billion per year (McGuire, 2018). There is also growing evidence in cyberpsychology showing that a significant number of people frequently engage in dishonest behaviors online, many of whom are arguably likely to go on and engage in uncooperative behaviors sometimes later in the future. For instance, a study by Caspi and Gorsky (2006) has found that one-third of a sample of web users reported to have engaged in online deception. More recently, Drouin et al. (2016) have demonstrated that almost 70% of their sample of 272 U.S. adults have lied at least sometimes to other people online. While not all uncooperative behaviors are dishonest (although all dishonest behaviors are arguably uncooperative by nature), statistics on dishonest behaviors seem to suggest that social media might in fact have an overall detrimental impact on human cooperation.

What explains the incongruity between previous evidence of people's capacity to maintain a high level of socially beneficial cooperation with strangers and the prevalence of dishonest

[1] One notable exception where expectation on contribution did not improve one's tendency to cooperate is in the TV version of the Golden Balls game, which is the game we have adapted for our lab and online experiments (Van den Assem, Van Dolder & Thaler, 2012).

behaviors on social media? One possibility is that while most social media users are generally cooperative citizens, there is nevertheless a small group of inherently "bad" individuals who frequently engage in large-scale dishonest behaviors in real-life as well as online. Without the ability to identify and exclude these individuals from using social media, it seems that there is little that SMIS providers could do to prevent them from behaving uncooperatively online. Another possibility is that people's preferences for cooperation are context-dependent, which could be improved or made worse in certain environments than others. If this was the case, then it becomes possible for SMIS providers to adjust rules and settings that would allow them to curb such an undesirable behavior. But currently the empirical evidence in this area is scarce, and the causal roots as well as the extent of people's preferences for cooperation on social media remain imperfectly understood.

This paper proposes a new empirical test of these theories. Through a series of randomized lab and online experiments, we tested whether the ability to hide one's real identity behind a fake profile – a social media-related behavior commonly known as 'catfishing' (Drouin et al., 2016; Marett et al., 2017) – is one of the main causes of uncooperative behaviors on social media websites as well as other online venues. By randomly allowing people to misrepresent one tiny information about themselves to others, i.e., their gender, we demonstrated that the average cooperation level with real money stake for the entire group was approximately 10-14 percentage points lower compared to other groups where the random opportunity to misrepresent was not available to the participants. This result suggests that social media, which makes it easy for people to misrepresent some information about themselves to others, is likely to have done more harm than good to human cooperation. Because of randomization, we can also conclude that cooperation is context-dependent in that people who would be cooperative in one setting behaved uncooperatively in another setting where there was a costless

opportunity to misrepresent their identity to others. We later discussed the economic impacts of having been randomized into such a scenario and whether we can predict who are more likely to misrepresent when given a chance. We finished by discussing the potential implications of our findings for SMIS providers and users.

This paper is outlined as follows. Section 2 briefly discusses the background literature. Section 3 describes the experimental design and the data. Empirical methods are outlined in Section 4. Results are then laid out in Section 5. Section 6 provides our discussions, the implications of our findings, and the concluding remarks.

2. Background and Hypotheses

There is a large literature devoted to understanding why people cooperate much more than predicted by standard economic models that assume rational and self-interest behaviors (see, e.g., Fehr & Fischbacher, 2002, and DellaVigna, 2009, for extensive reviews). One of the leading theories on cooperation is the theory of indirect reciprocity, which explains people's preferences for cooperation as a result of wanting to build trust and reputation that are essential in long-term interactions (Leimar & Hammerstein, 2001; Nowak & Sigmund, 2005). People may also cooperate because they fear altruistic punishment, that is, when individuals in a group incurs a cost to punish free-riders for not pulling in their weight in cooperative endeavors (Fehr & Gächter, 2000, 2002; Hermann, Thöni & Gächter, 2008; Rand & Nowak, 2013). Other studies have found evidence that individuals' preferences for cooperation may have stemmed from early in life (Olsen & Spelke, 2008) or are driven by personal characteristics that are close to being fixed over time (Vugt, Cremer & Janssen, 2007; Charness & Rustichini, 2011; Volk, Thöni & Ruigrok, 2011).

Many papers in economics have also shown that people's preferences for cooperation vary significantly across contexts. For example, allowing strangers to communicate with one another has been demonstrated to significantly increase cooperation in both one-shot and repeated social dilemma games (Miller, Butts & Rode, 2002; Bochet, Page & Putterman, 2006; Balliet, 2010). Other contextual effects on cooperation also include differences in cultural norms (Gächter, Hermann & Thöni, 2010), a disclosure of others' contributions (Frey & Meier, 2004), variations in the stake's size (Van den Assem, Van Dolder, & Thaler, 2012), and introducing subtle cues such as priming (Drouvelis, Metcalf & Powdthavee, 2015).

Modern technology like the internet and social media has dramatically changed the context under which people communicate and collaborate. However, there has been little empirical investigation into how these recent contextual changes affect people's preferences for cooperation. This study contributes to this small but increasingly relevant literature by focusing on the potential implications of social media (and other anonymized online venues in general) on human cooperation and economic well-being.

We argue that social media, which freely allows people to hide their real identity behind a fake profile – a social media-related behavior commonly known as 'catfishing' (Drouin et al., 2016; Marett et al., 2017) – substantially reduces human cooperation even when communication between people is allowed. We based this prediction on several theories in psychology and economics, as well as related evidence in social media research. First, we predict that the ability to misrepresent one's identity reduces one's fear of altruistic punishment (Fehr & Gächter, 2000, 2002; Hermann, Thöni & Gächter, 2008; Rand & Nowak, 2013) as social media enables them to move from one fake profile to another. Second, based on findings in social psychology

(Shu, Gino & Bazerman, 2011), the decision to engage in dishonest behavior of misrepresentation increases moral disengagement and motivates forgetting of the cooperative and honesty norms, which reduces feelings of guilt associated with uncooperative behaviors in the future. Third, the phenomenological theory of criminal spin (Ronel, 2010, 2011) suggests that the decision to misrepresent one's identity may start as something small and innocent, without malicious intent, but then escalates out of control into something bigger and significantly more damaging to others. Hence, the decision to catfish someone acts as a trigger that sets the flywheel of ever-intensifying antisocial behavior in motion. Forth, social media studies suggest that people derive a sense of enjoyment from engaging in online deception, thus making them more likely to lie as well as behave much more dishonestly later when communicating online than face-to-face (Naquin, Kurtzberg & Belkin, 2010; Caspi & Gorsky, 2006). Finally, people may choose to misrepresent some information about their identity for purely strategic reasons. For example, they might choose to misrepresent themselves as someone who shares interpersonal similarities with others to gain sympathy (Batson et al., 1995; Sally, 2001) before going on to defect later. Together, these theories help form two of this paper's main hypotheses:

Hypothesis 1: Uncooperative behavior on social media is context-dependent, i.e., there is a *cause-and-effect* that runs from giving people an opportunity to misrepresent their identity to reduced aggregate cooperation level and the economic welfare of the cybercrime victims. In other words, there will be a significant proportion of generally cooperative individuals who will go on to cheat others out of their money if they are given a chance to hide behind a fake online persona.

Hypothesis 2: By removing the opportunity to misrepresent one's identity online, we can substantially reduce people's tendency to engage in uncooperative behavior on social media and other online venues.

Evidence in support of hypothesis 1 implies that social media hinders rather than enhances human cooperation, whilst evidence in support of hypothesis 2 suggests that SMIS providers could do more in their system design to curb uncooperative behaviors between strangers on their social networking platforms. We test the above hypotheses as well as other welfare implications of misrepresentation opportunity through a series of lab and online experiments outlined in the next section.

3. Experimental Design

We recruited a total of 966 subjects – 686 students and 280 online participants in Prolific (www.prolific.co) – to participate in a one-shot variant of the prisoner's dilemma experiment with real money stakes. The experiment on the student sample was carried out in Warwick Business School's laboratory in June 2019 and January 2020, and in Nanyang Technological University (BTU) in Singapore in August 2019 and January 2020. We then carried out the same experiment with the same treatments online on the Prolific sample in April 2020. We were able to recruit a near gender balance sample of students for the lab experiment: 50.8% and 51.1% were male participants in Warwick and NTU, respectively. However, we were not able to do the same for the Prolific sample, ending up with 35.4% male participants for the online experiment.

In our experiment, which we preregistered through the Open Science Framework (OSF; https://osf.io/5q4hv), we randomized participants into the following four treatments:

1. Blind (or control)
2. True gender
3. Randomly assigned opportunity to misrepresent gender
4. Randomly assigned gender

We ran each treatment in random sessions; there were 29 lab sessions and one online session in total. In our lab experiments, once the student sample had arrived at the lab, they were randomly assigned to different numbered seats with partitions. On the other hand, participants in the online experiments were randomly paired using the oTree (https://www.otree.org) program as soon as they were recruited through Prolific. Once the experiment has started, all participants had to complete a questionnaire about their socio-demographic status, including age, gender, and – for the student sample – what they were majoring at the university.

In all treatments, two randomly matched players played a game called the "Golden Balls" game, which we adopted from a popular game show on TV in the UK and the Netherlands (Van den Assem, Van Dolder & Thaler, 2012; Turmunkh, Van den Assem & Van Dolder, 2019). Each player had to make an independent decision of whether to 'split' or 'steal' the money in a pot. If both players cooperated to split, they each received £10. If one chose to steal and the other chose to split, the person who stole received £20, and the person who cooperated received nothing. However, if both players stole, then both received nothing. Figure 1 displays the payoff matrix of the Golden Balls game that we showed the participants.

Figure 1: Golden Balls game payoff matrix

		Player B	
		Split	Steal
Player A	Split	£10, £10	£0, £20
	Steal	£20, 0	£0, £0

The game is a variant of the prisoner's dilemma with one crucial difference: Before each player made the steal or split decision, we allowed them 2 minutes to verbally communicate via an online messenger, similar to that of Facebook messenger, about what choice they planned to make as a pair. Both players were not allowed to identify themselves to each other, and that any agreement that they made in the chat would be non-binding, unverifiable, and costless if people want to go back on their word. We showed in Figure 1A in the Appendix that the most frequently used word during the chat across all treatments and within each treatment was "split", which suggests that cooperation is the intended signal that most people sent to each other, regardless of whether or not it was followed through. As can be seen in the Appendix, the word "split" is also easily the most frequently used word across all four treatments.

We gave players either true or false information about the other member's gender in all but the blind treatment ($N = 144$), in which each player played the game without any information about the other member. Players in the true gender treatment ($N = 224$) were told about the true gender of the other member in the pair before they had to make the split or steal decision. Although we consider the blind treatment as our control group, we also consider the true gender treatment as an alternative reference group as it represents a scenario where each player holds some real information about each other's identity.

To represent a social media scenario where users have an opportunity to misrepresent themselves to others, there were three conditions within the randomly assigned opportunity to misrepresent treatment ($N = 332$). For players in this treatment, one-half ($N = 166$) were randomly assigned an opportunity to misrepresent their gender to the other member. They had to make this decision after finding out the other member's gender. The other half were told that the other player was allowed to misrepresent their gender but may or may not take that opportunity to do so. Of those who were given the opportunity, 27% ($N = 45$) took it and mispresented their gender to the other player, while 72% ($N = 121$) received the opportunity but chose not to misrepresent.

We also tested whether randomly assigned gender – rather than randomly assigned opportunity to misrepresent gender – affects later cooperation. In other words, we wanted to test whether randomly forcing people to misrepresent their gender had made them more likely to steal in the Golden Ball games. There were three available conditions ($N = 266$) in this randomly assigned gender treatment. For these players, half ($N = 133$) were shown a randomly assigned gender of the other player that may or may not match their real gender. Of those who were randomly assigned gender, 47% ($N = 63$) knew that their gender was being randomly misrepresented, while 53% ($N = 70$) knew that their true gender was being shown to the other player.

Once participants finished making their split or steal decision, they were told of their earning from the Golden Balls game. After that, they were asked to complete a battery of post-experiment questions. This includes questions on subjective well-being, trust, risk preferences and personality.[2]

[2] For the description of the variables collected post-experiment, see online Appendix 2A.

While people also misrepresent other characteristics such as ethnicity and age online, one of the main reasons why we chose gender for this experiment was because it is by far the most frequently misrepresented information on social media (Drouin et al., 2016). It was also more practical to recruit students who identified themselves as either male or female for the lab experiment. There is also a wealth of research on gender differences in social dilemma experiments where we can draw upon to compare our results (Frank, Gilovich & Regan, 1993; Ortmann & Tichy, 1999; Croson & Gneezy, 2009).

Table 1: Selected characteristics by location of the experiment

Variables	SG-Lab	UK-Lab	UK/US-Online	Overall
Steal (=1)	0.36	0.43	0.18	0.33
	(0.03)	(0.03)	(0.02)	(0.02)
Payment	8.68	8.17	9.79	8.82
	(0.40)	(0.44)	(0.35)	(0.23)
Male	0.51	0.51	0.35	0.47
	(0.03)	(0.03)	(0.03)	(0.02)
Age	22.18	20.96	29.79	23.96
	(0.10)	(0.20)	(0.57)	(0.22)
Economics as major	0.08	0.12	N/A	0.07
	(0.01)	(0.02)		(0.01)
N	348	338	280	966

Note: The mean standard errors are in parentheses.

Table 1 displays some descriptive statistics by location of the experiment. On average, the steal rates were higher amongst participants in the lab experiments than in the online experiment. However, this could be due partly to the fact that there were more female as well as older participants in the online experiment than in the lab experiments. In all three locations, the average payment was slightly less than £10, which is the fair outcome in the Golden Balls game.

4. Empirical Methods

Our main analysis involves modeling the decision to split or steal in the Golden Balls game as a linear function of experimental conditions and personal characteristics, which can be written as follows

$$S_i = M_i'\beta + X_i'\gamma + \varepsilon_i. \quad (1)$$

Here, Eq. (1) assumes that individual *i* has a latent propensity to choose either split or steal S_i^*. However, we do not observe this latent variable, but the actual split or steal decision S_i, where $S_i = 0$ if the person chooses to split and $S_i = 1$ if the person chooses to steal. We impose the observation criterion $S_i = 1(S_i^* > 0)$, where $1(.)$ is the indicator function taking the value of 1 if $S_i^* > 0$ and 0 otherwise. The vector M_i' represents dummy variables representing different treatments and conditions within-treatment; X_i' indicates personal characteristics; and ε_i is the random error term. We estimated Eq. (1) using a binary probit model. However, given that probit coefficients are hard to interpret, the estimated marginal effects are reported instead in the results section. We also allowed for the possibility that the decisions of participants within the same experimental session are correlated by clustering the standard errors at the sessional level (see, e.g., Fréchette, 2012; Abadie, Athey, Imbens & Wooldridge, 2017).

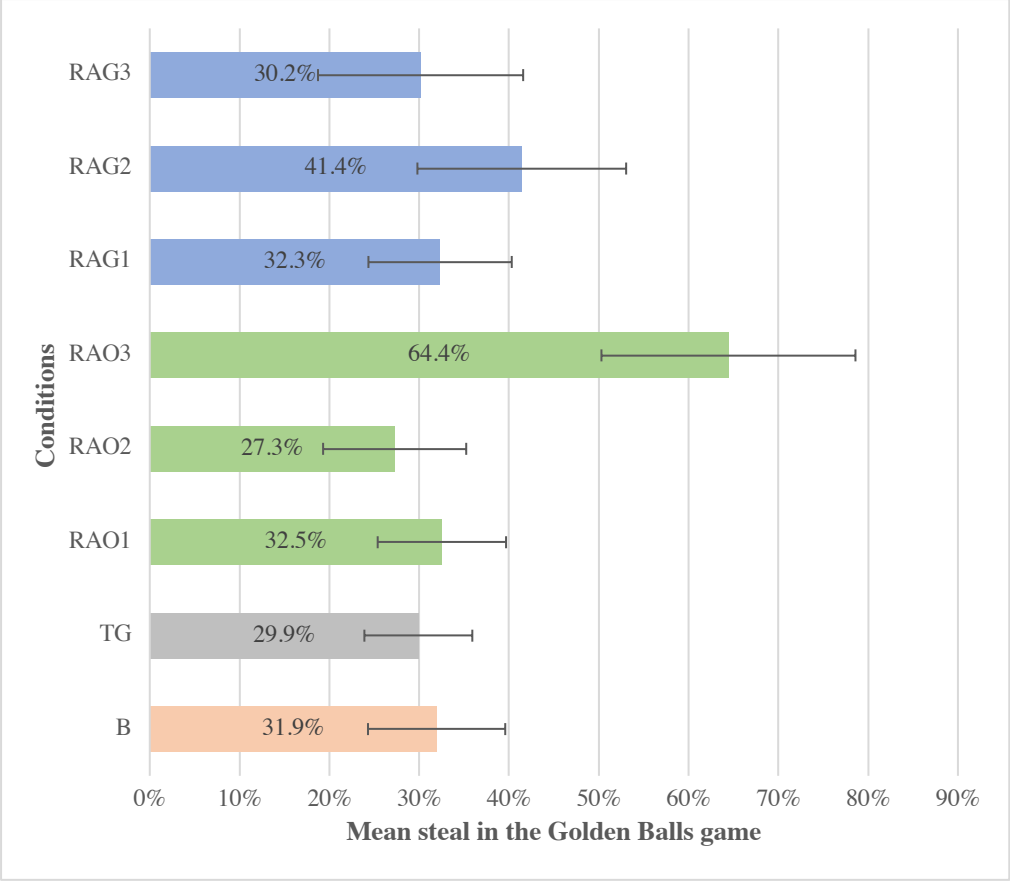
Other than analyzing how cooperation varies across treatments and conditions within treatment, we also tested whether income inequality was worse in certain contexts than others. We also carried out an analysis to predict the likelihood of misrepresentation, i.e., who are more likely to misrepresent? Finally, we carried out a simple before-and-after analysis of different emotions following a split or steal decision to learn how individual's subjective well-being was affected by different decisions.

5. Results

5.1. Cooperative Behavior: Split versus Steal

Figure 2 displays the raw data of mean steal decision across treatments and conditions within treatment. We found that people were generally cooperative – i.e., the average steal rate was less than 50% – in all conditions except for one. The average steal rate ranged from 27.3% in the randomly assigned opportunity to misrepresent but chose not to do so condition to 41.4% in the randomly assigned gender that matched their true gender condition. For those who were randomly assigned an opportunity to misrepresent and took it, the average steal rate was 64.4%, which was more than 30 percentage points higher than the average steal rates in the blind (31.9%) and the true gender (29.9%) treatment. A Wilcoxon signed rank test of steal in the misrepresentation group versus the blind treatment produced $z = -3.88, p = .000$ (two-tailed). The same test for the misrepresentation group versus the true gender group produced $z = -4.40, p = .000$ (two-tailed).

Figure 2: Mean steal decision in the Golden Balls game



Note: B = Blind treatment (control group); TG = True gender treatment; RAO1 = Randomly assigned opportunity to misrepresent treatment, condition#1: Did not receive opportunity to misrepresent gender; RAO2 = Randomly assigned opportunity to misrepresent treatment, condition#2: Randomly assigned opportunity, did not misrepresent; RAO3 = Randomly assigned opportunity to misrepresent treatment, condition#3: Misrepresented gender; RAG1 = Randomly assigned gender treatment, condition#1: Were not randomly assigned gender; RAG2 = Randomly assigned gender treatment, condition#2: Randomly assigned gender/matched own gender; RAG3 = Randomly assigned gender treatment, condition#3: Randomly assigned gender/mismatched own gender. Standard errors represent 95% confidence intervals.

As can be seen from Table 1, entering these treatments into a probit regression with control variables did not substantially change the overall findings. Controlling only for the demographic characteristics, Columns 1 and 2 showed that people who misrepresented their gender after having been randomly allowed to do so were 31.7 percentage points (95% CI: 23.3-40.1) and 36.6 percentage points (95% CI: 27.8-45.5) more likely to steal than those in the blind treatment and the true gender treatment, respectively. Due to randomization, it is

perhaps not surprising that Columns 3 and 4's results are also robust to controlling for individual's attitudes towards risks, the Dark Triad personality traits, and the general trust level.

Table 2: Marginal effects from probit regression on the decision to steal

VARIABLES	(1)	(2)	(3)	(4)
True gender treatment	-0.0459	REF	-0.00704	REF
	(0.0351)		(0.0286)	
Blind treatment	REF	0.0477	REF	0.00709
		(0.0379)		(0.0290)
Randomly assigned opportunity to misrepresent gender treatment				
i) Did not receive opportunity to misrepresent	0.0555*	0.105***	0.0992***	0.107***
	(0.0319)	(0.0347)	(0.0296)	(0.0305)
ii) Randomly assigned opportunity, did not misrepresent	0.0136	0.0619	0.0428	0.0502
	(0.0545)	(0.0691)	(0.0464)	(0.0516)
iii) Misrepresented gender	0.317***	0.366***	0.322***	0.330***
	(0.0429)	(0.0452)	(0.0515)	(0.0596)
Randomly assigned gender treatment				
i) Were not randomly assigned gender	0.0269	0.0757**	0.0921**	0.0998**
	(0.0420)	(0.0364)	(0.0455)	(0.0407)
ii) Randomly assigned gender/matched	0.109	0.161*	0.156**	0.164**
	(0.0782)	(0.0881)	(0.0696)	(0.0770)
iii) Randomly assigned gender/mismatched	0.0167	0.0654	0.0629	0.0706
	(0.0789)	(0.0868)	(0.0826)	(0.0900)
Personal characteristics				
Female matched with male	-0.165***	-0.165***	-0.112**	-0.112**
	(0.0408)	(0.0408)	(0.0445)	(0.0445)
Male matched with female	-0.0150	-0.0150	0.0106	0.0106
	(0.0416)	(0.0416)	(0.0406)	(0.0406)
Both females	-0.123***	-0.123***	-0.0794**	-0.0794**
	(0.0390)	(0.0390)	(0.0375)	(0.0375)
Age	-0.0451***	-0.0451***	-0.0384***	-0.0384***
	(0.0113)	(0.0113)	(0.00821)	(0.00821)
Age-squared	0.00059***	0.00059***	0.00051***	0.00051***
	(0.000134)	(0.000134)	(0.000103)	(0.000103)
Take Economics as major (if student)	0.0705	0.0705	0.0879	0.0879
	(0.0560)	(0.0560)	(0.0651)	(0.0651)
Singaporean sample	-0.0437	-0.0437	-0.0473*	-0.0473*
	(0.0283)	(0.0283)	(0.0286)	(0.0286)
Prolific (UK and US) sample	0.613***	0.613***	0.693***	0.693***
	(0.104)	(0.104)	(0.106)	(0.106)
Time taken in the prisoner's dilemma	-0.0099***	-0.0099***	-0.0107***	-0.0107***
	(0.00164)	(0.00164)	(0.00189)	(0.00189)
Risk taking attitudes			0.00784	0.00784
			(0.00674)	(0.00674)
Dark triad component: Narcissism			-0.00600	-0.00600
			(0.0185)	(0.0185)
Dark triad component: Psychopathy			0.0560***	0.0560***
			(0.0166)	(0.0166)
Dark triad component: Machiavellianism			0.142***	0.142***
			(0.0184)	(0.0184)
General trust			-0.263***	-0.263***
			(0.0271)	(0.0271)
Log pseudolikelihood	-545.06	-545.06	-472.36	-472.36

Observations	966	966	965	965

Note: *<10%; **<5%; ***<1%. Clustered-corrected standard errors at the sessional level are reported in parentheses. Dependent variable is a binary variable: 0 = split, 1 = steal. The marginal effects are estimated at the means. Note that one person in the online sample got logged out before completing the post-questionnaire.

We also found evidence that the possibility of being paired with someone who might be misrepresenting their gender, whether by choice or by chance, significantly heightened the probability that the individual will go on to steal as well. Compared to the blind and the true gender treatments, people who did not receive the opportunity to misrepresent in the randomly assigned opportunity to misrepresent gender treatment were approximately 10 percentage points more likely to steal in the Golden Balls game. The same applies to those who were not randomly assigned a gender in the exogenous gender treatment. These findings suggest that the uncertainty of being catfished also substantially lowered lower cooperation amongst individuals who were not handed an opportunity to misrepresent as well. Only those who were randomly allowed to misrepresent but then refused to do so were consistently honest in their later behavior as those participating in the blind and the true gender treatments.

If misrepresentation causes uncooperative behavior, then we should also expect to see people who were randomly assigned a gender that mismatched their own to defect more often than those in the blind and the true gender treatments. However, there was little evidence to suggest that they did. By contrast, we found some evidence that people who were randomly assigned a gender that matched their own were approximately 16 percentage points more likely to steal compared to those in the blind and the true gender treatments. Both of these findings are counter-intuitive and harder to interpret. One possibility is that forced misrepresentation may have led individuals who did not have the intention to misrepresent to overcompensate by becoming more cooperative than they would have otherwise. On the other hand, there was no reason for individuals who were randomly assigned a gender that matched their own to feel

guilty about misrepresentation. However, unlike participants in the randomly assigned opportunity to misrepresent, they were also not allowed the choice to be honest about their gender either. This may have eliminated the reinforcement opportunity that comes from a recent honest behavior of choosing to be honest about one's gender. In addition to this, the possibility of being thought of as a catfish by the other player may have increased the probability of stealing as well. Nevertheless, these explanations are mere speculations and a further investigation into people's motivations is required to explain these results.

Perhaps a more crucial question than whether people who misrepresented are more likely to defect is whether randomly allowing people to misrepresent their gender had made the whole group behave significantly worse, on average. To answer this question, Table 3 reported the intention-to-treat (ITT) (Hollis & Campbell, 1999) effects of the randomly assigned opportunity to misrepresent and the randomly assigned gender treatments on the decision to steal. By including both compliers and non-compliers in the estimation, the ITT effects of randomly assigned opportunity to misrepresent gender were 12.2 percentage points (95% CI: 0.03-21.4) when compared to the blind treatment, and 13 percentage points (95% CI: 0.03-22.7) when compared to the true gender treatment. On the other hand, the ITT effects of randomly assigned gender were 11.1 percentage points (95% CI: 0.002-22.0) when compared to the blind treatment, and 12 percentage points (95% CI: -0.004-24.5) when compared to the true gender treatment. Qualitatively similar results were also obtained if we were to include those who were not given the opportunity to misrepresent as well. These estimates imply that, on average, the entire participants in both treatments behaved significantly worse than individuals in treatments where there was either zero or real information about the other player's gender identity.

Table 3: The intention-to-treat effect on the decision to steal (marginal effects probit estimator)

VARIABLES	(1)	(2)
True gender treatment	-0.00807	REF
	(0.0283)	
Blind treatment	REF	0.00813
		(0.0288)
Randomly assigned opportunity to misrepresent gender treatment		
i) Did not receive opportunity to misrepresent	0.101***	0.110***
	(0.0295)	(0.0309)
ii) ITT effect of randomly assigned opportunity to misrepresent	0.121**	0.130***
	(0.0473)	(0.0494)
Randomly assigned gender treatment		
i) Were not randomly assigned gender	0.0925**	0.101**
	(0.0455)	(0.0407)
ii) ITT effect of randomly assigned gender	0.111**	0.120*
	(0.0554)	(0.0638)
Implied treatment effects		
Randomly assigned opportunity to misrepresent gender treatment	0.109***	0.117***
	(0.0293)	(0.0312)
Randomly assigned gender treatment	0.101***	0.109***
	(0.0359)	(0.0396)
Other control variables as in Columns 3 and 4 in Table 1	Yes	Yes
Log pseudolikelihood	-476.99	-476.99
Observations	965	965

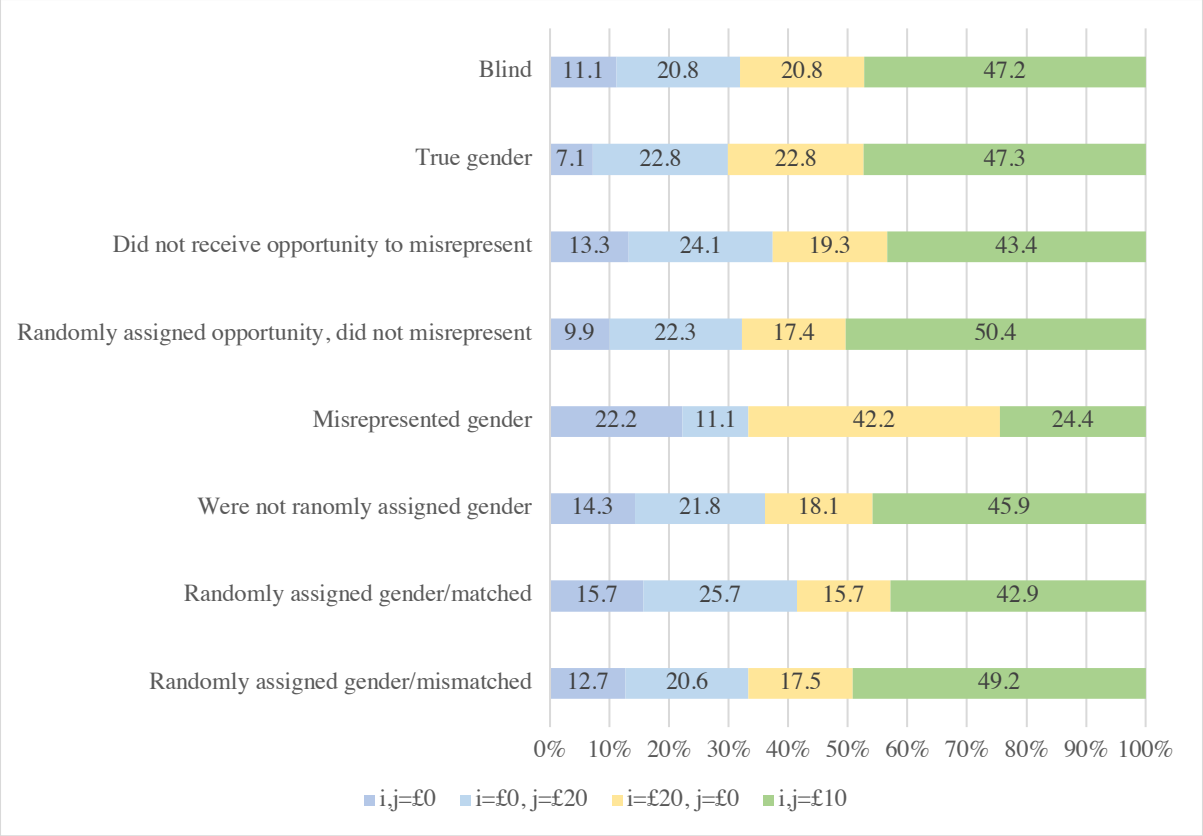
Note: *<10%; **<5%; ***<1%. Clustered-corrected standard errors at the sessional level are reported in parentheses. Dependent variable is a binary variable: 0 = split, 1 = steal. The marginal effects are estimated at the means. Note that one person in the online sample got logged out before completing the post-questionnaire.

What about the misrepresentation effects themselves? How significantly damaging were they to human cooperation? To get some idea, we compared them to the gender effect, as illustrated in Figure 3A in the Appendix. Consistent with previous findings on gender differences in the Golden Balls game on TV (Van den Assem, Van Dolder & Thaler, 2012) and the prisoner's dilemma game played in a lab (Frank, Gilovich & Regan, 1993; Ortmann & Tichy, 1999; Croson & Gneezy, 2009), we found men in the blind treatment were 10.4 percentage points (95% CI: 0.04-17.2) more likely to steal than women in the same treatment. However, this gender disparity disappeared amongst women who chose to misrepresent their gender as they were 31 percentage points (95% CI: 0.03-59.0) more likely to steal, on average.

5.2. Income inequality and economic welfare

An important welfare question is whether participants who were randomly assigned to play against catfishers suffered an extreme financial loss from participating in the prisoner's dilemma game. Figure 2, which displays the income distribution across experimental conditions, shows a considerable income disparity between catfishers and those who randomly got catfished. Amongst individuals who misrepresented their gender, 42% of people who defected went on to earn the maximum income (£20) from the Golden Balls game (*i*=steal, *j*=split). This is more than double the proportion of individuals who defected and received £20 in return in other conditions. Approximately 22% of catfishers earned nothing from defecting as the other player also acted uncooperatively (*i*=split, *j*=split), while 11% cooperated and lost all their earnings as a result of the other person choosing to defect in the game (*i*=split, *j*=steal). Only 24% cooperated and earned the fair payment (*i*=split, *j*=split) of £10 as a result of cooperation by both members in a pair. Altogether, roughly 64% of people who were randomly paired with a catfisher received £0 from the Golden Balls game. For all other conditions reported in Figure 2, we can see that income is much more equally distributed across participants in any given pair. Across all other conditions, more than 40% of people earned a fair payment of £10 each.

Fig 2. Distribution of individual earning by conditions



Note: The notation *i* represents the player assigned to that condition, and *j* represents the other player in the pair. Hence, *i*=£20, *j*=£0 implies that player *i* chose to steal and player *j* chose to split in the Golden Balls game.

We also applied the Gini coefficient, a standard measure of income inequality, to formally assess the extent of income disparity across different conditions. The Gini coefficient (G), which is defined as the mean relative difference between a uniformly random pair of observed values, has a value that ranges from 0 (strict equality) and 1 (maximum inequality).[3]

We found $G = 0.43$ for players in the blind treatment, $G = 0.42$ for players in the true gender treatment, $G = 0.41$ for players who were not randomly assigned the opportunity to misrepresent and those who were but did not misrepresent, $G = 0.57$ for players who were not randomly assigned the opportunity to misrepresent and those who were and decided to misrepresent, $G = 0.47$ for players who were not randomly assigned gender and those who

[3] We calculated the Gini coefficient by applying STATA's code *ineqdec0* on the payment variable for each condition (Jenkin, 2008).

were randomly assigned a gender that matched their own, and $G = 0.42$ for players who were not randomly assigned gender and those who were randomly assigned a gender that mismatched their own. These results indicate that income inequality is highest within pairs of catfishers and people who were randomly paired to play against them.

We also found the financial losses suffered by those who were randomly playing against a catfish to be enormous compared to other groups. Holding other characteristics constant, we can see from Table 4 that the average catfisher in the randomly assigned opportunity to misrepresent treatment was 17.3 (95% CI: -0.003-34.9) and 12.3 (95% CI: -0.07-31.4) percentage points more likely earn maximum payment from the Golden Balls game than those in the blind and the true gender treatments. Conversely, the results also imply that people who were randomly paired with a catfisher were between 12-17 percentage points more likely to walk away from the game with nothing.

Table 4: Marginal effects from probit regression on earning maximum (£20)

VARIABLES	**(1)**	**(2)**
True gender treatment	0.0388	REF
	(0.0273)	
Blind treatment	REF	-0.0358
		(0.0232)
Randomly assigned opportunity to misrepresent gender treatment		
i) Did not receive opportunity to misrepresent	0.0233	-0.0145
	(0.0335)	(0.0330)
ii) Randomly assigned opportunity, did not misrepresent	-0.000338	-0.0360
	(0.0306)	(0.0315)
iii) Misrepresented gender	0.173**	0.123
	(0.0898)	(0.0973)
Randomly assigned gender treatment		
i) Were not randomly assigned gender	0.0101	-0.0266
	(0.0441)	(0.0351)
ii) Randomly assigned gender/matched	0.0768**	0.0337
	(0.0390)	(0.0475)
iii) Randomly assigned gender/mismatched	0.00382	-0.0320
	(0.0549)	(0.0557)
Other control variables as in Columns 3 and 4 in Table 1	Yes	Yes
Log pseudolikelihood	-476.99	-476.99
Observations	965	965

Note: *<10%; **<5%; ***<1%. Clustered-corrected standard errors at the sessional level are reported in parentheses. Dependent variable is a binary variable: 0 = earned less than £20, 1 = earned maximum £20. The marginal effects are estimated at the means. Note that one person in the online sample got logged out before completing the post-questionnaire.

We also carried out several robustness checks. First, despite running the same experiment across different contexts, we found the results to be qualitatively similar across locations (UK, Singapore, and the USA), as well as across genders. For example, we showed in Table 5 that the steal rate was highest among those who took the opportunity to misrepresent themselves in the randomly assigned opportunity treatment in all contexts. The estimated marginal effects were similar in size and statistically significantly different from zero in all but one, i.e., the online condition. Table 6's results illustrated that men and women who misrepresented their gender were equally more likely than the blind group to steal in the Golden Balls game. These results provide reassurance that the main results were not driven by a particular subset of environment.

Table 5: Marginal effects from probit regression on the decision to steal by environments

	Lab			Online
VARIABLES	**UK/SG-lab**	**UK-lab**	**Singapore-lab**	**UK/USA-online**
True gender treatment	0.0131	-0.0187	0.0221	-0.0788
	(0.0296)	(0.0366)	(0.0405)	(0.0513)
Randomly assigned opportunity to misrepresent gender treatment				
i) Did not receive opportunity to misrepresent	0.115**	0.0608	0.149**	0.0209
	(0.0447)	(0.0683)	(0.0608)	(0.0677)
ii) Randomly assigned opportunity, did not misrepresent	0.0466	0.0580	-0.000190	-0.0127
	(0.0819)	(0.136)	(0.0837)	(0.0637)
iii) Misrepresented gender	0.322***	0.413***	0.253***	0.274
	(0.0607)	(0.108)	(0.0609)	(0.225)
Randomly assigned gender treatment				
i) Were not randomly assigned gender	0.133***	0.218***	0.0575	-0.0384
	(0.0488)	(0.0719)	(0.0406)	(0.0721)
ii) Randomly assigned gender/matched	0.141*	0.216	0.100*	0.179
	(0.0817)	(0.141)	(0.0568)	(0.165)
iii) Randomly assigned gender/mismatched	0.0255	-0.148	0.177*	0.0872
	(0.112)	(0.142)	(0.0939)	(0.131)
Personal characteristics				
Female matched with male	-0.131**	-0.152*	-0.101	-0.0503
	(0.0611)	(0.0901)	(0.0969)	(0.0636)
Male matched with female	-0.00930	-0.0742	0.0484	0.0442
	(0.0523)	(0.0805)	(0.0666)	(0.0798)

Both females	-0.101**	-0.186***	-0.0106	-0.0428
	(0.0512)	(0.0710)	(0.0575)	(0.0695)
Age	0.0590	0.0589	0.114	-0.0234**
	(0.0509)	(0.0860)	(0.106)	(0.0113)
Age-squared	-0.00169	-0.00200	-0.00244	0.000320**
	(0.00105)	(0.00189)	(0.00223)	(0.000153)
Take Economics as major (if student)	0.0913	0.0448	0.211**	N/A
	(0.0707)	(0.107)	(0.0990)	
Time it took to make the decision in the prisoner's dilemma	-0.0117***	-0.0111***	-0.0109***	N/A
	(0.00209)	(0.00263)	(0.00336)	
Risk taking tendency (0=risk averse, …, 10=risk-loving)	0.0105	0.00280	0.0157	0.00412
	(0.00986)	(0.0119)	(0.0153)	(0.00951)
Risk taking attitudes	0.0131	0.00340	0.0320	-0.0377
	(0.0206)	(0.0405)	(0.0266)	(0.0246)
Dark triad component: Narcissism	0.0533**	0.0533*	0.0610	0.0468*
	(0.0222)	(0.0307)	(0.0385)	(0.0262)
Dark triad component: Psychopathy	0.157***	0.136***	0.173***	0.0810***
	(0.0274)	(0.0460)	(0.0289)	(0.0273)
Dark triad component: Machiavellianism	-0.273***	-0.341***	-0.217***	-0.197***
	(0.0373)	(0.0405)	(0.0540)	(0.0411)
General trust	-0.0117***	-0.152*	-0.101	-0.0503
	(0.00209)	(0.0901)	(0.0969)	(0.0636)
Log pseudolikelihood	-363.87	-174.16	-180.11	-102.46
Observations	686	338	348	279

Note: *<10%; **<5%; ***<1%. Clustered-corrected standard errors at the sessional level are reported in parentheses. Dependent variable is a binary variable: 0 = split, 1 = steal. The marginal effects are estimated at the means. Location dummy is included as a control in Column 1. Note that one person in the online sample got logged out before completing the post-questionnaire.

Table 6: Marginal effects from probit regression on the decision to steal by gender

VARIABLES	Men	Women
True gender treatment	-0.0308	0.0387
	(0.0353)	(0.0571)
Randomly assigned opportunity to misrepresent gender treatment		
i) Did not receive opportunity to misrepresent	0.0954*	0.100
	(0.0566)	(0.0764)
ii) Randomly assigned opportunity, did not misrepresent	0.0791	-0.0112
	(0.0548)	(0.0598)
iii) Misrepresented gender	0.323**	0.345***
	(0.136)	(0.121)
Randomly assigned gender treatment		
i) Were not randomly assigned gender	0.0453	0.161**
	(0.0560)	(0.0776)
ii) Randomly assigned gender/matched	0.192**	0.0924
	(0.0860)	(0.111)
iii) Randomly assigned gender/mismatched	0.113	0.00356
	(0.0796)	(0.126)
Other control variables as in Columns 3 and 4 in Table 1	Yes	Yes
Log pseudolikelihood	-233.44	-233.59
Observations	517	448

Note: *<10%; **<5%; ***<1%. Clustered-corrected standard errors at the sessional level are reported in parentheses. Dependent variable is a binary variable: 0 = split, 1 = steal. The marginal effects are estimated at the means.

We also demonstrated that the misrepresentation effect on the decision to steal was far more substantial for people who were randomly paired with someone of the same gender; see Figure 4A in the Appendix. This last finding suggests that people who were matched with someone of the same gender may choose to strategically misrepresent their gender to maximize their payoff instead of the joint payoff. Finally, we showed in Appendix 5A that people who received £0 reported the largest drop in the positive emotions factor, which is defined by changes in feelings of enthusiasm, interest, and excitement. On the other hand, those who received £10 ("split, split" decisions) reported the largest increase in the same factor. A Wilcoxon signed rank test of the positive emotions factor between people who received £0 and £10 produced $z = -19.93, p = .000$ (two-tailed). The same test for people who received £10 and £20 produced $z = 14.41, p = .000$ (two-tailed). Perhaps surprisingly, we found people who received £20 to report the most significant increase in the negative emotions factor that is defined by changes in feelings of fear, nervousness, scared, and guilt. Here, a Wilcoxon signed rank test of the pre-post changes in the negative emotions factor between people who received £20 and £10 produced $z = -9.64, p = .000$ (two-tailed). The same test for people who received £0 and £10 produced $z = -4.77, p = .000$ (two-tailed). While there were some differences in people's responses to life satisfaction questions, the differences by payment were not statistically significantly different from zero.

5.3. Predicting the decision to misrepresent

One question of interest is whether we can reasonably target who is likely to misrepresent their gender identity based on their observable characteristics. We examined this question in Table 7 by estimating a probit regression with the decision to misrepresent as the dependent variable on a sample of individuals who were randomly assigned an opportunity to misrepresent ($N = 166$). Apart from the location variable, in which the Prolific sample was the least likely to

misrepresent, we found that none of the observable characteristics such as gender, age, and whether or not the individual is studying for an economics degree strongly predict the decision to misrepresent. This result suggests that it may be difficult for policymakers to target specific groups of individuals who may be more likely to create fake profiles on social media based on their observable characteristics alone. However, by adding psychological traits that we collected but otherwise generally unobservable into the regression model, we found some evidence Narcissism and Psychopathy to marginally predict the decision to misrepresent.

Table 7: Marginal effects from probit regression on the decision to misrepresent

VARIABLES	**(1)**	**(2)**
Male	0.0978	0.0382
	(0.0932)	(0.0973)
Age	-0.0165	0.00515
	(0.0233)	(0.0191)
Age-squared	0.000291	2.94e-05
	(0.000288)	(0.000235)
Take Economics as major (if student)	-0.0180	-0.000491
	(0.109)	(0.0954)
Singaporean sample	0.0822	0.0708
	(0.101)	(0.117)
Prolific (UK and US) sample	-0.176**	-0.222***
	(0.0784)	(0.0773)
General risk preference		0.0285
		(0.0182)
Dark triad component: Narcissism		0.0759*
		(0.0418)
Dark triad component: Psychopathy		0.0581*
		(0.0321)
Dark triad component: Machiavellianism		0.0438
		(0.0414)
Trust		-0.0733
		(0.0861)
Log pseudolikelihood	-90.03	-84.17
Observations	166	166

Note: *<10%; ***<1%. Clustered-corrected standard errors at the sessional level are reported in parentheses. The sample consists of those who were randomly allowed the opportunity to misrepresent. Dependent variable is a binary variable: 0 = received an opportunity to misrepresent but did not take it, 1 = misrepresented. The marginal effects are estimated at the means.

5.4. Constraints on generality

In this subsection, we take the opportunity to express what we believe to be the constraints on the generality of our findings (Simons et al., 2017). We have shown that a significant proportion of participants who chose to misrepresent their gender went on to behave uncooperatively in a prisoner's dilemma. The findings were robust across different locations and cultural settings. Thus, we expect the results to generalize across samples taken from countries other than Singapore, UK, and USA. While we do not have evidence that the findings will be reproducible for other types of strategic games such as trust game, public goods game, and dictator game, we believe that our results would have been generalizable in settings where there is a randomized opportunity for people to misrepresent. Given that most if not all of our student participants use social media on a daily basis, we also believe the results will be reproducible with a sample of randomly selected social media users across different countries. It also remains to be seen whether our results can be generalized to scenarios where the stakes are large as well. Finally, we have no reason to believe that the results depend on other characteristics of the subjects, materials, or context.

6. Discussion and Conclusion

Previous research has shown that strangers are often capable of maintaining a high level of socially beneficial cooperation, even when communication is not permissive between individuals. It is, therefore, not entirely unreasonable to assume that social media, which revolutionized human connection, is, in principle, beneficial to human cooperation as well. However, our results suggest that allowing people even small short-term opportunities to misrepresent one's identity, which is a common feature of most social media platforms, was enough to corrode this cooperative norm and, in turn, causes a substantial reduction in later cooperation as well as significant pecuniary and nonpecuniary damages for those who are

naturally cooperative. Nevertheless, misrepresentation was not necessarily the cause of later uncooperative behavior; people who were misrepresenting their gender by the luck of the draw were, on average, no more likely to defect than those in the blind and the true gender treatments. Instead, it is the combination of an opportunity and the intention to misrepresent that damages human cooperation. In addition to this, we also found some evidence that the fear of "being catfished" and "being perceived as a catfish by others" had lowered the average cooperation for the entire group as well.

Our findings have significant policy implications for the prevention of cybercrimes. According to one Facebook's estimate[4], 5% of monthly active accounts are fake, which is equivalent to around 13 million Facebook profiles. They also have a very selective policy on authentication, e.g., only 'public figures' who have a significant number of followers and a media presence are eligible to have their Facebook page verified. Given what we know from our 'true gender' treatment, in which players with limited though truthful information about the identity of other players were able to maintain high levels of socially beneficial cooperation, allowing people to verify their social media page seems to be crucial to combating future online frauds. Such a policy would allow the opportunity to misrepresent to be 'decoupled' (Gladwell, 2019) from people's intention to commit cybercrime, which, we hope, should produce a similar deterrence effect on undesirable behaviors as other decoupling decisions in the past (Seiden, 1978). For example, by progressively removing carbon monoxide from the public gas supply, policymakers in England and Wales managed to decouple the ready access to a means of death from the intention to commit suicide, which led to a significant drop in the number of suicides between 1965 and 1975 (Clark & Mayhew, 1988). We believe that an increase in the number

[4] Facebook -- An Update on How We Are Doing at Enforcing Our Community Standards [website]. 2020. Retrieved from https://about.fb.com/news/2019/05/enforcing-our-community-standards-3/

of verified users will drive out individuals with fake profiles, which should, in turn, increase cooperation among strangers whose profiles have already been authenticated. Future research may have to come back to assess the impact of a large-scale enrolment of opportunities to verify on cooperation among social media users.

References

Abadie, A., Athey, S., Imbens, G.W. and Wooldridge, J., 2017. *When should you adjust standard errors for clustering?* (No. w24003). National Bureau of Economic Research.

Balliet, D., 2010. Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, *54*(1), pp.39-57.

Batson, C.D., Batson, J.G., Todd, R.M., Brummett, B.H., Shaw, L.L. and Aldeguer, C.M., 1995. Empathy and the collective good: Caring for one of the others in a social dilemma. *Journal of Personality and Social Psychology*, *68*(4), p.619.

Bochet, O., Page, T. and Putterman, L., 2006. Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, *60*(1), pp.11-26.

Caspi, A. and Gorsky, P., 2006. Online deception: Prevalence, motivation, and emotion. *CyberPsychology & Behavior*, *9*(1), pp.54-59.

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

Charness, G. and Rustichini, A., 2011. Gender differences in cooperation with group membership. *Games and Economic Behavior*, *72*(1), pp.77-85.

Clarke, R.V. and Mayhew, P., 1988. The British gas suicide story and its criminological implications. *Crime and justice*, *10*, pp.79-116.

Croson, R. and Gneezy, U., 2009. Gender differences in preferences. *Journal of Economic Literature*, *47*(2), pp.448-74.

DellaVigna, S., 2009. Psychology and economics: Evidence from the field. *Journal of Economic Literature*, *47*(2), pp.315-72.

Drouin, M., Miller, D., Wehle, S.M. and Hernandez, E., 2016. Why do people lie online? "Because everyone lies on the internet". *Computers in Human Behavior*, *64*, pp.134-142.

Drouvelis, M., Metcalfe, R. and Powdthavee, N., 2015. Can priming cooperation increase public good contributions? *Theory and Decision*, *79*(3), pp.479-492.

Fischbacher, U., Gächter, S. and Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics letters*, *71*(3), pp.397-404.

Fehr, E. and Fischbacher, U., 2002. Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. *Economic Journal*, *112*(478), pp.C1-C33.

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

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

Frank, R.H., Gilovich, T. and Regan, D.T., 1993. Does studying economics inhibit cooperation? *Journal of Economic Perspectives*, *7*(2), pp.159-171.

Fréchette, G.R., 2012. Session-effects in the laboratory. *Experimental Economics*, *15*(3), pp.485-498.

Frey, B.S. and Meier, S., 2004. Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review*, *94*(5), pp.1717-1722.

FTC Press Release [website]. 2020. Retrieved from https://www.ftc.gov/news-events/press-releases/2020/02/new-ftc-data-show-consumers-reported-losing-more-200-million

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

Gächter, S., Herrmann, B. and Thöni, C., 2010. Culture and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), pp.2651-2661.

Gladwell, M., 2019. *Talking to strangers: What we should know about the people we don't know*. Penguin UK.

Herrmann, B., Thöni, C. and Gächter, S., 2008. Antisocial punishment across societies. *Science*, *319*(5868), pp.1362-1367.

Hollis, S. and Campbell, F., 1999. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*, *319*(7211), pp.670-674.

Jenkins, S.P., 2008, December. Estimation and interpretation of measures of inequality, poverty, and social welfare using Stata. In *North American Stata Users' Group Meetings 2006*(No. 16). Stata Users Group.

Leimar, O. and Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1468), pp.745-753.

Marett, K., George, J.F., Lewis, C.C., Gupta, M. and Giordano, G., 2017. Beware the dark side: Cultural preferences for lying online. *Computers in Human Behavior*, *75*, pp.834-844.

McGuire, M., 2018. Into the web of profit. *Understanding the Growth of Cybercrime Economy. Bromium*.

Miller, J.H., Butts, C.T. and Rode, D., 2002. Communication and cooperation. *Journal of Economic Behavior & Organization*, *47*(2), pp.179-195.

Naquin, C.E., Kurtzberg, T.R. and Belkin, L.Y., 2010. The finer points of lying online: E-mail versus pen and paper. *Journal of Applied Psychology*, *95*(2), p.387.

Nowak, M.A. and Sigmund, K., 2005. Evolution of indirect reciprocity. *Nature*, *437*(7063), pp.1291-1298.

Olson, K.R. and Spelke, E.S., 2008. Foundations of cooperation in young children. *Cognition*, *108*(1), pp.222-231.

Ortmann, A. and Tichy, L.K., 1999. Gender differences in the laboratory: evidence from prisoner's dilemma games. *Journal of Economic Behavior & Organization*, *39*(3), pp.327-339.

Paulhus, D.L. and Williams, K.M., 2002. The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), pp.556-563.

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

Ronel, N. 2010. Criminal behavior, criminal mind: Being caught in a criminal spin. *International Journal of Offender Therapy and Comparative Criminology*. doi:10.1177/0306624X10384946

Ronel, N., 2011. Criminal behavior, criminal mind: Being caught in a "criminal spin". *International Journal of Offender Therapy and Comparative Criminology*, *55*(8), pp.1208-1233.

Sally, D., 2001. On sympathy and games. *Journal of Economic Behavior & Organization*, *44*(1), pp.1-30.

Seiden, R.H., 1978. Where are they now? A follow-up study of suicide attempters from the Golden Gate Bridge. *Suicide and Life-Threatening Behavior*, *8*(4), pp.203-216.

Simons, D. J., Shoda, Y., & Lindsay, D. S., 2017. Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123-1128.

Shu, L.L., Gino, F. and Bazerman, M.H., 2011. Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin*, *37*(3), pp.330-349.

Turmunkh, U., van den Assem, M.J. and Van Dolder, D., 2019. Malleable lies: Communication and cooperation in a high stakes TV game show. *Management Science*, *65*(10), pp.4795-4812.

Van den Assem, M.J., Van Dolder, D. and Thaler, R.H., 2012. Split or steal? Cooperative behavior when the stakes are large. *Management Science*, *58*(1), pp.2-20.

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

Vugt, M.V., Cremer, D.D. and Janssen, D.P., 2007. Gender differences in cooperation and competition: The male-warrior hypothesis. *Psychological Science*, *18*(1), pp.19-23.

Online Appendix [Not to be published]

Fig 1A: Text analysis of the 2-minute communication across all treatments

Note: Text analysis on the 2-minute communication between players across all treatments shows that "split" is the most frequently used words in the conversation. The larger the words are in this word cloud, the more frequent they appeared in the conversation.

Blind (Control)

True Gender

Randomly assigned opportunity to misrepresent

Randomly assigned gender

Appendix 2A: Description of the variables collected in the post-experiment questionnaire

Well-being measures included positive and negative emotions (Watson, Clark & Tellegen, 1988) and life satisfaction scales (Diner et al., 1985). We measured positive and negative emotions twice, once before and once after players played the Golden Balls game. There were twenty questions on the individual's usual state of positive and negative emotions: interested, distressed, excited, upset, strong, guilty, scared, hostile, enthusiastic, proud, irritable, alert, ashamed, inspired, nervous, determined, attentive, jittery, active, and afraid. We then generated the change variables for each of these emotions, e.g., change in upset = post-upset *minus* pre-upset, before conducting a factor analysis to derive with two factors of emotions: positive and negative.

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Uniqueness
change_int~d	0.6878	0.0816	-0.1721	-0.0030	-0.0242	-0.0694	0.0518	-0.0210	0.4821
change_dis~d	-0.0995	0.2862	0.3460	0.1328	-0.0195	-0.1058	0.2698	0.0066	0.6864
change_exc~d	0.7587	0.0781	-0.2035	0.0105	-0.0412	-0.0378	0.0056	0.0242	0.3731
change_upset	-0.4101	0.0960	0.6078	0.1670	0.0018	-0.0571	0.0996	-0.0164	0.4118
change_str~g	0.5345	-0.0558	0.0709	0.0167	0.0408	0.0948	-0.0658	-0.0455	0.6888
change_gui~y	0.0148	0.3716	0.0588	0.6486	-0.0292	-0.0160	0.0097	-0.0165	0.4360
change_sca~d	0.0283	0.7273	0.1069	0.2004	-0.0701	0.0331	-0.0478	-0.0568	0.4071
change_hos~e	-0.1895	0.2023	0.4602	0.1489	-0.0152	0.1051	-0.1579	-0.0672	0.6485
change_ent~c	0.7237	-0.0193	-0.2224	-0.0415	0.0635	0.0385	-0.0283	0.0164	0.4181
change_proud	0.5064	-0.2040	-0.0692	-0.1175	0.0278	0.2869	-0.0248	0.0218	0.5991
change_irr~e	-0.3162	0.1353	0.6222	0.0421	-0.0149	-0.0193	-0.0383	0.0410	0.4891
change_alert	0.2881	0.2313	0.1397	-0.0494	0.2249	-0.0219	-0.0998	0.0164	0.7803
change_ash~d	-0.0869	0.3256	0.1600	0.6348	0.0204	-0.0110	0.0078	0.0218	0.4567
change_ins~d	0.4807	-0.0840	-0.1284	-0.0364	0.1531	0.3337	-0.0586	-0.0262	0.6052
change_ner~s	0.0531	0.6649	0.0123	0.1238	0.0970	-0.1057	0.0603	0.0296	0.5145
change_det~d	0.5143	0.0793	-0.0449	-0.0953	0.2772	0.1060	-0.0129	-0.0720	0.6247
change_jit~y	0.1197	0.4487	0.0982	0.2091	0.1571	-0.0644	0.1203	0.1471	0.6661
change_act~e	0.5300	0.0745	-0.0998	0.0180	0.3345	0.1189	0.0196	0.0431	0.5750
change_afr~d	-0.0147	0.7288	0.1057	0.1864	0.0194	0.0050	0.0088	0.0197	0.4219

We also carried out the same factor analysis for the five questions on life satisfaction: life is close to ideal (sat1), excellent conditions of life (sat2), satisfied with life (sat3), got important things in life (sat4), and change nothing about life (sat5).

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
sat1	0.7922	0.1465	0.3510
sat2	0.7873	0.1093	0.3682
sat3	0.8390	0.2032	0.2547
sat4	0.7193	0.2251	0.4319
sat5	0.5750	0.2185	0.6217

Trust was measured by asking participants the following question: "Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with other people?". Participants responded with either "You can't be too careful" or "Most people can be trusted".

Risk preferences were evaluated by asking participants how willing they were to take risks on a scale from 1 to 10, where low scores represent risk aversion. Risk preferences were also measured for specific contexts (i.e., while driving, with your health, in your occupation) using the same 1 to 10 scale.

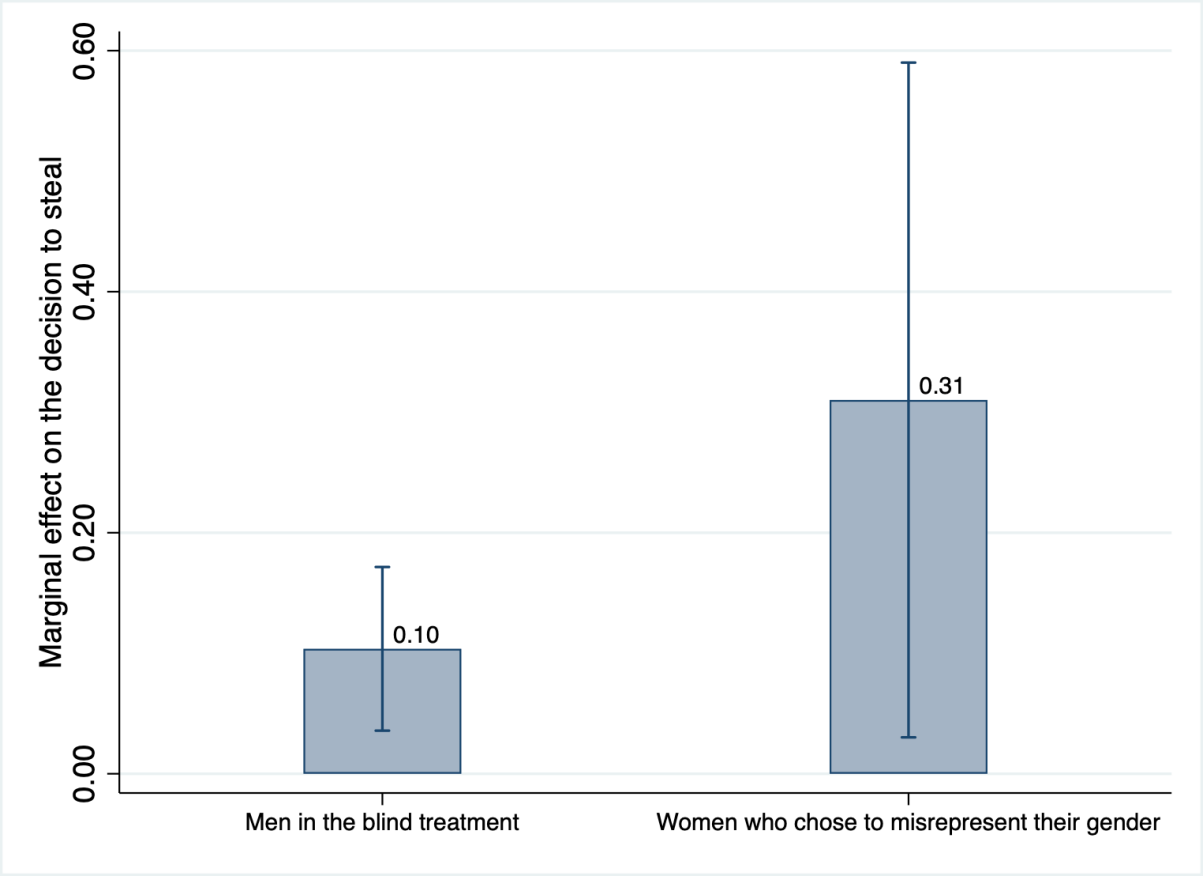
Participants also filled out the "Dirty Dozen" Dark Triad personality questionnaire to measure narcissism, psychopathy and Machiavellianism (Paulhus & Williams, 2002). We then conducted a factor analysis on the twelve variables to derive with three main factor components of the dark personality traits.

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness
personality1	0.2519	0.4159	0.6484	-0.1145	0.0922	0.3216
personality2	0.1744	0.2780	0.7352	0.0570	-0.0363	0.3471
personality3	0.2777	0.1511	0.6357	0.1148	-0.0909	0.4745
personality4	0.2155	0.4976	0.6361	-0.0416	0.0745	0.2941
personality5	0.1515	0.6991	0.3404	-0.0572	-0.0283	0.3684
personality6	0.1119	0.6785	0.2841	-0.0191	-0.0362	0.4448
personality7	0.1462	0.7098	0.2530	0.0868	0.0486	0.4010
personality8	0.1763	0.4481	0.2908	0.2584	-0.0013	0.6168
personality9	0.8365	0.0445	0.1339	0.1132	-0.0167	0.2673
personali~10	0.8507	0.0901	0.1389	-0.0070	-0.0151	0.2486
personali~11	0.7450	0.2102	0.2253	-0.0717	0.0373	0.3435
personali~12	0.5998	0.3078	0.2148	-0.1604	0.0632	0.4696

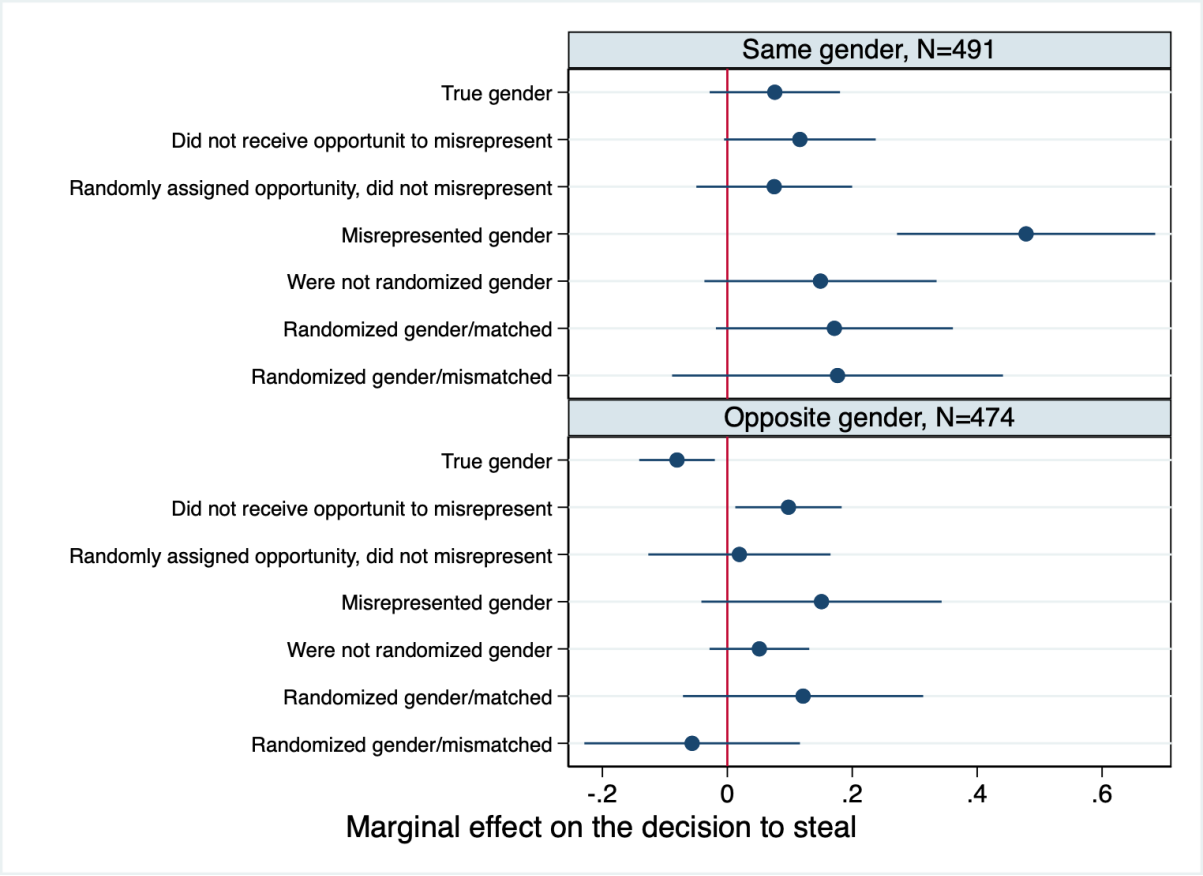
For the lab experiment, the participants left at the end of the experiment with the money they earned from the Golden Balls game + £2 participation fee. For the online experiment, participants were paid the money they earned from the Golden Balls game + $2 participation fee (£1.55).

Fig 3A: Comparing the marginal effects of men in the blind treatment versus women who chose to misrepresent in the randomly assigned opportunity to misrepresent treatment



The reference group is women in the blind treatment. The standard-error bars represent 95% confidence intervals.

Fig. 4A: The marginal effects on the decision to steal in the Golden Balls game by gender pairing



Note: The estimated marginal effects are based on a Probit regression with steal as the outcome variable. The regression controlled for age, age-squared, gender pairing, the time it took to make the split or steal decision, and a dummy representing whether the subject is currently studying economics (if the subject is a student). The horizontal line represents the baseline, i.e., the blind treatment. The standard-error bars represent 95% confidence intervals.

Appendix 5A: Summary of positive, negative, and life satisfaction by payment categories

```
. tab payment, su(emotion1)

           |  Summary of Scores for factor 1
   payment |      Mean   Std. Dev.       Freq.
-----------+------------------------------------
         0 | -.73882103    .846172         320
        10 |  .6532739   .57565898         440
        20 | -.24765915   .69672845         206
-----------+------------------------------------
     Total | -2.309e-11   .93768718         966

.
. tab payment, su(emotion2)

           |  Summary of Scores for factor 2
   payment |      Mean   Std. Dev.       Freq.
-----------+------------------------------------
         0 | -.34381789   1.0201915         320
        10 | -.04681871   .70186886         440
        20 |  .63408716   .81794318         206
-----------+------------------------------------
     Total |  9.180e-10   .91450111         966

.
. tab payment, su(psat)

           |  Summary of Scores for factor 1
   payment |      Mean   Std. Dev.       Freq.
-----------+------------------------------------
         0 | -.04641741   .98711698         320
        10 |  .03837837   .90079224         440
        20 |  -.0098685   .92977655         206
-----------+------------------------------------
     Total |  1.231e-09   .93610747         966
```

Note: Emotion1 = positive emotions factor, which is defined by changes in feelings of enthusiasm, interest, and excitement. Emotion2 = negative emotions factor, which is defined by changes in feelings of afraid, nervousness, scared, and guilt. Psat = life satisfaction.

Online Appendix References

1. Watson, D., Clark, L. a., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

2. Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75.

3. Paulhus, D.L. and Williams, K.M., 2002. The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), pp.556-563.

Experimental Instructions

[Note: These are the instructions that subjects faced in the blind (or control) treatment. Adjustments for the 'True gender', 'Randomly assigned opportunity to misrepresent gender', and 'Randomly assigned gender' treatments are shown in square brackets.]

Instructions

This experiment consists of 2 parts. In part 1, you will be asked to fill out a short demographic survey. You will then be given a £10 endowment.

Next, you will be randomly paired with another person in the room and given instructions on how to play the Golden Ball game (see attached). You will then be given two minutes to negotiate your decision with the other person using an instant messenger system. Please note that whatever decisions you and the other person may have been discussing during the allocated two minutes are not binding, and you can still make any decision you want when prompted with a decision later. After two minutes have passed, the instant messaging program will end, and you will no longer be able to talk to the other person. You will then be asked for your decision (steal/split, this will be explained in full below). Once both participants have submitted their decision, both will be shown the outcome.

In part 2, you will fill out a short 10 minute questionnaire.

Restrictions on messages

1. You must not identify yourself or send any information that could be used to identify you (for example, your name, contact details or seat in the room);
2. You must not make any threats, insults or use any obscene or offensive language.

If you violate these rules your payment will be forfeited.

You will be paid in private with cash at the end of the experiment.

Please raise your hand if you have any questions at any point during the experiment.

GOLDEN BALLS GAME INSTRUCTIONS

You have been randomly paired with another person in the lab. You will not learn the identity of the other person, during or after today's session, [*True gender/Randomly assigned opportunity to misrepresent gender/Randomly assigned gender treatment:* but you will be given information about the gender of the other person.] Both you and the other person have been given £10 to play Golden Ball which will be added to a communal pot (£20) to be won. You and the other person will then be given two minutes to use an instant messaging program to discuss what will happen to this money. You can choose to split the money or steal the money.

If **you choose split**, and the **other person chooses split**, you both get £10 in addition to your participation fee (£12 each).
If **you choose steal**, and the **other person chooses steal**, you both get £0 in addition to your participation fee (£2 each).
If **you choose steal**, and the **other person chooses split**, you get the full pot of money (£20) in addition to your participation fee (£22), and the other person gets £0 in addition to the participation fee (£2).
If **you choose split**, and the **other person chooses steal**, you get £0 in addition to your participation fee (£2), and the other person gets £20 in addition to the participation fee (£22).

During these two minutes, you can negotiate with the other person as to whether you plan to split the money and try to change their mind if they plan to steal. Messages will be shared *only* between you and the other person. While you are free to say whatever you want in this portion of the experiment, you are not allowed to identify yourself and that abusive, threatening or offensive language will not be tolerated. These messages will be read later to check for language of this type (participants found engaging in abusive, threatening or offensive language will be report and subject to disciplinary action). After two minutes have passed, the instant messaging program will end and you will no longer be able to talk to the other person. You will then be given a choice to steal or split the money. Please note that you are free to make any decision you want to make, which may be the same or completely different to what you and the other person have agreed upon during the two minutes chat. After both participants have made a decision, your decision will be shared with the other person.

[*Randomly assigned opportunity to misrepresent gender treatment:* **One randomly selected participant from a pair will be given the opportunity to misrepresent their gender to the other person. This means that a participant who is male will be offered the option to misrepresent themselves as female and a participant who is female will be offered the option to misrepresent themselves as male. If you are given the option to misrepresent your gender, the other person will NOT be given the option to misrepresent their gender but will be aware that you were given an option to misrepresent your gender to him/her. If allocated to this condition, you can choose whether you want to misrepresent your gender or not.**]

[*Randomly assigned gender treatment*: **One randomly selected participant from a pair will be assigned a gender randomly determined by the computer. There is a 50% chance that the randomly selected participant will assigned to be male and a 50% chance that the randomly selected participant will be assigned to be female. If you were randomly assigned to be male, your opponent will be told you are male regardless of your true gender, and vice versa if you were randomly assigned female. If you are the randomly**

selected participant whose gender will be determined by the computer at random, your opponent will NOT be given the option to misrepresent their gender to you, but will be aware that you were randomly assigned a gender, which may or may not be representative of your true gender identity.]