Titel der Arbeit:

Dishonesty: The role of rewards, professional identity and experimenter purpose disclosures

(Unehrlichkeit: Die Rolle von Belohnungen, beruflicher Identität und der Offenlegung des Versuchszwecks)

Dissertation
zur Erlangung des akademischen Grades Doctor of Philosophy
(Ph.D.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
Zoe Rahwan (McLaren)

Präsident (komm.)
der Humboldt-Universität zu Berlin

Prof. Dr. Peter Frensch

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

Prof. Dr. Dr. Christian Ulrichs

Gutachter/innen

1. **Prof. Dr. Ralph Hertwig**
2. **Prof. Dr. Bernd Irlenbusch**
3. **Prof. Dr. Urs Fischbacher**

Tag der mündlichen Prüfung: 25. April 2022

For my girls, Norah and Miriam

Acknowledgements: I would like to thank Dan Ariely for inspiring me, through his online course 'A Beginner's Guide to Irrational Behavior,' to return to studies after many years in industry and for provoking my interest in researching honesty. I am grateful to Robert Dixon, my Honours Economics supervisor from Melbourne University, who has continued to show an interest in my work over the years and to Peter Bossaerts who made an extraordinary accommodation to support my further studies.

My gratitude is extended to my former regulatory bosses Peter Casey, Peter Smith and Ian Johnston who always provoked my intellectual curiosity at work, always pushed me to refine my thinking and opinions, and supported my studies at LSE. I would like to thank all of my lecturers at LSE, including Jeroen Nieboer who took time to guide the design of my first experiment, Heather Barry-Kappes who continues to support my research to this day, and Barbara Fasolo - my Masters supervisor and on-going collaborator, who effortlessly coordinated the often complicated administration of various projects. I am grateful for the support of my other collaborators Ewa Kochanowska, Oliver Hauser and Erez Yoeli. Thanks to Jonathan Schulz and Moshe Hoffman who saw value in a collection of null results and encouraged me to publish them.

Thank you to Jenn Lerner and the Lerner Lab at Harvard Kennedy School for providing an academic home for me while in Cambridge and opening my eyes to the role of emotions in decision making. My thanks to Ralph Hertwig for his nudge (boost?) to run an experiment (Project 3), members of the Center Adaptive Rationality for providing an incredible environment to conduct research and to the extraordinary administrators across MPIB who provided support. Thanks to Jürgen Rossbach for his cheerful, professional graphics assistance.

All of this research could literally not have been done without the participation of curious and courageous banking and regulatory field partners, their staff, and thousands of Amazon Mechanical Turk workers, and a Qualtrics panel, cleverly arranged by James Gorman, and of course funding from Google ATAP, arranged by Galen Pickard. My gratitude to the editors at Nature and the Journal of Economic Behavior and Organization, and the reviewers who did an extraordinary job mentoring me from a distance. My deepest thanks to my Humboldt University PhD Committee members of whom I am honoured to name as, Jule Specht (Committee Chair), Ralph Hertwig, Bernd Irlenbusch (also a reviewer for Project 1), and Urs Fischbacher.

Finally, my thanks go to friends and family for their love and support (including those who teased me for the phenomena which I study - thanks Jo and Graham Jones). Thank you to my parents for always encouraging my interest in learning, my mum for proof-reading and to all my grandparents who inspired me with their varied interests and who helped me understand that formal education is not necessarily the key to understanding the world. Thanks to Anna Uebergang who patiently listens to me talking about research and who proofread the thesis. My gratitude to my lovely neighbours, Klaus and Monika Danzer, who provided translation assistance and helped me adjust to a neighbourhood and city. A special thankyou to Kristin Augsburg, my dear long-time friend, who both incentivised and helped me move to her beloved Berlin and endures tales of research and my terrible German. Much gratitude is extended to Charley Wu and Christina Leuker for their assistance in navigating administrative labyrinthes. My deep gratitude goes to my husband, Iyad Rahwan, whose enthusiasm for my research I cannot curb. His support, and that from Arifa Awadallah, Debbie Menon, Emma O'Brien, Heloise Widdig, Annie Stingl, Constance McKenna, Ngaio Mango-Thurgood, Elmi Amran, Keatyn Jane and Jes Walsh together helped to keep our household joyful, rested and healthy. And last but not least, my thanks to Chilly Gonzales and Lambert for providing the soundtrack for this work.

Abstract

When and why do people decide to behave dishonestly? By understanding dishonest behaviour, policy makers are better able to deter such behaviour and to support a thriving society and economy. The study of dishonesty has flourished in recent years, driven by the establishment of crowd-sourced labour platforms, though some important field work has also emerged. The empirical findings from these studies have supported the emergence of new economic and psychological models to explain dishonest behaviour. Yet, how replicable and generalisable are leading experimental findings? And what other contextual factors -- like the nature of reward, scale of reward, and design choices from the experimenter-- may drive dishonest behaviour?

The central focus of this thesis was the attempted replication of a heavily cited paper in academia and the popular press. Previous replication efforts by-passed this work given the challenge of accessing professional participants. The paper which we attempted to replicate found that only bankers whose professional identity was made salient behaved dishonestly. This work was based on the notion that priming, or making salient one aspect of an individual's identity and the associated norms, would affect behaviour. As priming professional banking identity prompted dishonesty, this was concluded to be indicative of problematic norms in the banking sector. Though it was unclear if this finding would hold with other banks, for example in the same or other jurisdictions, in different segments (e.g. commercial versus investment banking), and over time.

In undertaking attempts to replicate the original study, I confronted a number of other empirically open questions, such as how variation in large-sized rewards, the nature of reward (i.e. for oneself or for charity) and the nature of stated experimental purpose (i.e. false purpose, incomplete disclosure, true purpose) affects dishonesty in commonly used experimental tasks. In doing so, I explored the generalisability of existing findings both with regard to the role of professional identity in generating dishonest behaviour and the sensitivity of dishonesty to rewards at a largely unexplored scale. Further, I generated novel insights on how the nature of reward affected honesty in a large-scale field study and on how the methodological consideration of stated experimental purpose can affect the measurement of dishonesty.

I explored these considerations across ten experiments. In Chapter 1, I outline the motivation for each of the projects, each of which are summarised in Chapter 2. Briefly, in Project 1, I executed four field experiments with difficult-to-access financial services professionals and one experiment with a panel. This work, assessing the role of priming financial services professional identities on honesty, was conducted in three regions around the world. In Project 2, I executed three experiments via a crowd-sourced labour platform lab to understand both the role of reward size in dishonesty and spillovers onto subsequent charitable giving and moral feelings (e.g. guilt). In Project 3, I executed two large-scale experiments on a crowd-sourced labour platform using two commonly used experimental tasks to understand how and why stated experimental disclosure may affect dishonest behaviour. In Chapter 3, I provide a general discussion of the findings, reflecting upon the limitations of the work and various caveats that apply. I also suggest some future research directions. The published manuscripts for Projects 1 and 2, and draft manuscripts for Project 3 are contained in Chapters 4, 5 and 6 respectively.

Zusammenfassung

Wann und warum entscheiden sich Menschen für unehrliches Verhalten? Durch das Verständnis von unehrlichem Verhalten sind politische Entscheidungsträger besser in der Lage, ein solches Verhalten zu verhindern und eine florierende Gesellschaft und Wirtschaft zu unterstützen. Das Studium der Unehrlichkeit hat in den letzten Jahren eine Blütezeit erlebt, angetrieben durch die Etablierung von Crowd-Sourced-Arbeitsplattformen, obwohl auch einige wichtige Feldarbeiten entstanden sind. Die empirischen Erkenntnisse aus diesen Studien haben die Entstehung neuer ökonomischer und psychologischer Modelle zur Erklärung unehrlichen Verhaltens unterstützt. Doch wie replizierbar und verallgemeinerbar sind die führenden experimentellen Ergebnisse? Und welche anderen kontextuellen Faktoren wie die Art und das Ausmaß der Belohnung und die Designentscheidungen des Experimentators können unehrliches Verhalten beeinflussen?

Im Mittelpunkt dieser Arbeit stand der Versuch der Replikation einer in der akademischen Welt und in der populären Presse viel zitierten Arbeit. Frühere Replikationsversuche haben diese Arbeit umgangen, da es schwierig war, Zugang zu professionellen Teilnehmern zu bekommen.
Die Arbeit, die wir zu wiederholen versuchten, ergab, dass nur Banker, deren berufliche Identität hervorgehoben wurde, sich unehrlich verhielten. Diese Arbeit basierte auf der Vorstellung, dass das Priming, also das Hervorheben eines Aspekts der Identität einer Person und der damit verbundenen Normen, das Verhalten beeinflussen würde. Da das Priming der professionellen Bankidentität Unehrlichkeit auslöste, wurde daraus geschlossen, dass dies ein Hinweis auf problematische Normen im Bankensektor ist. Es war jedoch unklar, ob dieses Ergebnis auch für andere Banken gilt, z. B. in der gleichen oder einer anderen Gerichtsbarkeit, in verschiedenen Segmenten (z. B. Commercial versus Investment Banking) und im Zeitverlauf.

Bei dem Versuch, die ursprüngliche Studie zu replizieren, habe ich mich mit einer Reihe anderer empirisch offener Fragen auseinandergesetzt, z.B. wie die Höhe der Belohnung bei großen Einsätzen, die Art der Belohnung (d.h. für sich selbst oder für einen wohltätigen Zweck) und die Art des angegebenen Versuchszwecks (d.h. falscher Zweck, unvollständige Offenlegung, wahrer Zweck) die Unehrlichkeit in häufig verwendeten Versuchsaufgaben beeinflusst. Dabei untersuchte ich die Verallgemeinerbarkeit bestehender Befunde sowohl in Bezug auf die Rolle der beruflichen Identität bei der Generierung von Unehrlichkeitsverhalten als auch auf die Empfindlichkeit von Unehrlichkeit gegenüber Belohnungen in weitgehend unerforschtem Ausmaß. Darüber hinaus habe ich neue Erkenntnisse darüber gewonnen, wie die Art der Belohnung die Ehrlichkeit in einer groß angelegten Feldstudie beeinflusst hat und wie die methodische Berücksichtigung des erklärten Versuchszwecks die Messung der Unehrlichkeit beeinflussen kann.

Ich habe diese Überlegungen in zehn Experimenten untersucht. In Kapitel 1 skizziere ich die Motivation für jedes der Projekte, die in Kapitel 2 zusammengefasst werden. Kurz gesagt, in Projekt 1 habe ich vier Feldexperimente mit schwer zugänglichen Finanzdienstleistern und ein Experiment mit einem Panel durchgeführt. Diese Arbeit, die die Rolle des Primings von professionellen Identitäten im Finanzdienstleistungsbereich auf Ehrlichkeit untersucht, wurde in drei Regionen auf der ganzen Welt durchgeführt. In Projekt 2 führte ich drei Experimente über ein Crowd-Sourced-Arbeitsplattform-Labor durch, um sowohl die Rolle der Belohnungsgröße bei Unehrlichkeit als auch die Spillover-Effekte auf nachfolgende wohltätige Spenden und moralische Gefühle (z.B. Schuld) zu verstehen. In Projekt 3 habe ich zwei groß angelegte Experimente auf einer Crowd-Sourced-Work-Plattform

durchgeführt, die zwei häufig verwendete experimentelle Aufgaben verwenden, um zu verstehen, wie und warum die angegebene experimentelle Offenlegung unehrliches Verhalten beeinflussen kann. In Kapitel 3 liefere ich eine allgemeine Diskussion der Ergebnisse und reflektiere die Grenzen der Arbeit und verschiedene Vorbehalte, die gelten. Ich schlage auch einige zukünftige Forschungsrichtungen vor. Die veröffentlichten Manuskripte für die Projekte 1 und 2 sowie die Manuskriptentwürfe für Projekt 3 sind in den Kapiteln 4, 5 bzw. 6 enthalten.

Chapter 1: Introduction

Background: The Replication Crisis

Honesty, among other pro-social behaviours, is fundamental to a flourishing society. The scientific understanding of when and why people behave honestly has rapidly expanded in recent years, predominantly via experiments in laboratories and on-line platforms (Gerlach, Teodorescu, and Hertwig 2019) though also via field work (Gächter and Schulz 2016; Cohn et al. 2019). Older, narrowly-focused economic-based cost-benefit models of behaviour (Becker 1968) are being challenged and augmented by models explicitly incorporating psychological aspects of behaviour (Akerlof George and Kranton Rachel 2000; Mazar, Amir, and Ariely 2008; Gneezy, Kajackaite, and Sobel 2018). Yet, some of the most recent influential studies regarding honesty have not themselves escaped the 'reproducibility crisis' (Alexander et al. 2012; Open Science Collaboration 2015; Pashler and Wagenmakers 2012; "Altmetric – The Dishonesty of Honest People: A Theory of Self-Concept Maintenance" 2020; Kristal et al. 2020) and other influential findings have been found to be based on fabricated data (Leif, Simonsohn, and Simmons 2021). The influential work of Cohn et al (2014), which assesses that professional culture can play in dishonesty, had evaded replication attempts due to the practical challenge of accessing hard-to-reach populations of bankers (Colin F. Camerer et al. 2018).

Highly-cited findings in the psychological and social sciences are being found frequently to fail to replicate at conventional levels of significance and have substantially smaller effect sizes in replication attempts (C. F. Camerer et al. 2016; Colin F. Camerer et al. 2018; Ioannidis 2008; Open Science Collaboration 2015). While replication efforts have gone far beyond one topic, the notion of priming was one of the earliest areas targeted for criticism of non-replication (Cesario 2014). Priming studies deploy cues which have a nonconscious influence on subsequent behaviour. Early non-replications were difficult to publish (Yong 2012) and in one particular case (Cohn, Fehr, and Marechal 2014), the ability to access populations impeded even an attempt at replication (Colin F. Camerer et al. 2018). In addition to the 'reproducibility crisis', the ability to generalize findings is commonly, but not always (Klein et al. 2018), undermined by limitations of the population studied (Henrich, Heine, and Norenzayan 2010).

These factors together raise significant challenges for the advancement of understanding human decision-making and have been argued to form the basis for a *scientific revolution* (Kuhn 2021). That is, as current experimental findings and their contradictory findings cannot be well explained within predominant frameworks, researchers should look to systematically study how sampling and moderators drive heterogeneity in effects (Bryan, Tipton, and Yeager 2021). This effort could also embrace the notion of representative design, in which researchers attempt to sample experimental stimuli and conditions in a representative manner (Brunswik 1955). Together, these efforts could assist in deepening scientific understanding, rebuilding public confidence in the experimental social sciences and provide greater utility (and avoid harm) of scientific findings in policy-making.

Professional Identity and Dishonesty

Cohn et al (2014) find that bankers, as opposed to other professionals, are more likely to behave dishonestly when their professional identity is made salient. The finding was motivated by a theoretical model which suggests that individuals have several identities (e.g. based on gender, profession, religion) and that by activating, or priming, one particular identity, this would evoke relevant norms and consequently affect behaviour (Akerlof George and Kranton Rachel 2000; Bargh, Chen, and Burrows 1996). The finding that priming banking identity led to dishonesty has been interpreted to suggest that the culture of banking, which features highly variable performance-based remuneration (Conrads et al. 2014), had a 'corrosive' effect on honesty (Mohan 2014).

This work, published in the wake of the 2008 global financial crisis and amid poor trust in bankers, garnered significant publicity ("Altmetric – Business Culture and Dishonesty in the Banking Industry" 2020). In academia specifically, it was published at a time of elevated concerns over lack of reproducibility of priming results (Cesario 2014) though still remains heavily cited (Serra-Garcia and Gneezy 2021). Given the difficulty of recruiting bankers – particularly in a study on honesty – one of the largest efforts at replicating recent high-profile behavioural science studies published in Science and Nature excluded this study (Colin F. Camerer et al. 2018).

In Project 1, the key objective was to explore whether the original findings would replicate across bankers and other non-banking professionals, in addition to assessing how professional identity would affect financial services regulators - those responsible for overseeing the banking and other financial sectors. By deploying the same task in different populations, sourced from different national jurisdictions (i.e. 5 samples in 3 jurisdictions), I aimed to understand the limits to generalising the original finding, including how variation in national culture could affect dishonesty (Gächter and Schulz 2016).

The Nature of Rewards and Dishonesty

There is a gap in the literature regarding how qualitative differences in the nature of a payoff available can affect dishonesty. That is, does it affect behaviour whether an individual's dishonesty leads to tangible benefits for themselves or others? This question however emerged as one of practical importance in our replication attempts. Specifically, we found that in running field experiments our banking field partners were content to use selfish rewards (i.e. shopping vouchers) for their staff though regulatory field partners were not. That is, the financial services regulators were unwilling to participate in a study which enabled their staff to engage in unethical behaviour that would lead to them being personally enriched. As such, we amended the nature of the reward to be a charitable reward when running these studies with regulators. For regulatory staff, their winnings would be awarded to a charity affiliated with their organisation.

The existing literature did not provide direct insights on how a different category of rewards, while holding the scale of reward constant, affects dishonesty. However, if one conceives of the charitable reward as being equivalent to a USD0 selfish reward, the existing literature (Fischbacher and Föllmi-Heusi 2013; Mazar, Amir, and Ariely 2008; Abeler, Nosenzo, and Raymond 2019) would suggest that participants would be largely insensitive to the size of the reward. It was also unclear if that literature would generalise to large scale rewards (maximum USD140) in the field, with USD50 marking the maximum reward in experimental tasks at the time (Kajackaite and Gneezy 2017).

Understanding how the nature of reward can affect honesty formed the second objective for Project 1. With one large sample of bankers we experimentally varied the nature of the reward, being either for oneself (i.e. a shopping voucher) or for others (i.e. a charitable donation made on their behalf) and assessed the effect on dishonesty.

The Size of Reward and Dishonesty

The cornerstone economic model of criminal behaviour, often adopted for immoral behaviour more broadly, suggests that individuals calculate the external costs, benefits and likelihood of being detected when taking a decision to behave against societal norms (Becker 1968). The model does not explicitly account for internal costs (and benefits) of engaging in immoral behaviour, such as cheating. Such internal costs may help to explain why Becker’s model is challenged by experimental findings that participants are insensitive to rewards (Fischbacher and Föllmi-Heusi 2013; Mazar, Amir, and Ariely 2008). Indeed, a meta-analysis of 90 honesty studies found that participants on average only claimed three quarters of the maximal payoff (Abeler, Nosenzo, and Raymond 2019). Why don’t people cheat to the maximal extent, particularly when there are seemingly negligible costs to doing so?

Mazar et al (Mazar, Amir, and Ariely 2008) proposed an alternative theory, that of *self-concept maintenance*, which introduces internal costs and benefits to acting dishonestly. In this model, individuals do seek to extract external rewards (e.g. monetary payoffs) from dishonest behaviour though this is only done to the extent where they can maintain a self-concept as being a moral person. Similarly, Fischbacher and Föllmi-Heusi (Fischbacher and Föllmi-Heusi 2013) propose that as the size of the external reward increases so too does the marginal cost of the lie, thereby deterring maximal cheating. Gneezy et al. (Gneezy, Kajackaite, and Sobel 2018) build, formally and experimentally, upon the idea of internal costs of dishonesty, proposing three types of internal or lying costs: “a cost related to the distance between the true outcome and what is reported; a cost related to the monetary gains generated by the lie; and a cost associated with the probability that a statement is perceived to be dishonest.” They conclude, supported by experimental evidence, that the perception of dishonesty dominates the other considerations. That is, the notion of one’s identity - which can be driven both by an intrinsic motivation to behave appropriately and a motivation to appear to be doing so - is more important than influences of the size of the payoff or probability of a favourable outcome.

By varying rewards by 500-fold, with an upper limit of a nearly unprecedented USD50, we aimed to both confirm that (i) models which included some internal cost of dishonesty outperformed traditional cost-benefit models (Becker 1968), (ii) that insensitivity of dishonesty to rewards prevailed at extreme levels of rewards, providing confidence that small variations in our studies in Project 1 compared with the original study (Cohn, Fehr, and Marechal 2014) would not account for differences in results and (iii) to measure how internal or psychological costs varied with rewards varying, using a measure of self-perceived morality, and the durability of any costs incurred by engaging in dishonesty. To assess the durability of any spillovers from engaging or refraining from dishonest behaviour, I took two measures of self-reported morality; one immediately after completing the honesty task, and one a day later. This enabled an assessment of how participants’ perception may be affected over time, including whether they engage in “unethical amnesia” (Kouchaki and Gino 2016).

Another objective was to understand how past dishonest behaviour can spillover on subsequent prosocial behaviour through the lens of psychological licensing (Miller and Effron 2010; Zhong, Liljenquist, and Cain 2009). Specifically, I explored whether dishonesty – or refraining from dishonesty – affects subsequent decisions regarding the propensity to donate, and for those opting to make a donation, the proportion of winnings to be donated. There is accumulating evidence that individuals can give themselves a psychological license to behave or express an opinion considered to be morally unacceptable having previously behaved or expressed in a morally acceptable manner (Miller and Effron 2010). In this way, individuals engage in some form of balancing their ‘good’ and ‘bad’ behaviour over time.

Experimental Purpose Disclosures and Dishonesty

False purpose deception is the most commonly used type of deception in psychology studies (Hertwig and Ortmann 2008). It involves a situation in which participants "may be given, or be caused to hold, false information about the main purpose of the study" from an experimenter (Sieber, Iannuzzo, and Rodriguez 1995). In contrast to psychologists, who broadly are tolerant of false purpose and other forms of deception, economists exhibit a strong aversion to the practice. This can be reflected in economic and other journals with economics editors desk-rejecting work using deception, while funding applications containing proposed use of deception may be similarly denied (McDermott 2013; Jamison, Karlan, and Schechter 2008; Cooper 2014; Cook and Yamagishi 2008).

The core of economists' concerns regarding the use of deception appears to be the belief that participants are 'tainted' by such treatment. That is, trust between an experimenter and participant, commonly viewed among economists as a public good, may be undermined when experimenters deliberately mislead participants. Following exposure to deception, there is an expectation that participants may not be able to fully rely upon information given by researchers and therefore may behave differently. This loss of experimental control would consequently invalidate tests of economic theory (Ariely and Norton 2007; Barrera and Simpson 2012; McDermott 2013).

By contrast, psychologists have a relatively long history of using deception in experiments. For psychologists, deception can be a means to study important aspects of the human condition. Without a means to deceive participants, particularly regarding the nature of the study, clean measures cannot be obtained. Psychologists argue that this could lead to a loss of collective knowledge about critical (and sometimes, uncomfortable) behaviours like obedience (Milgram 1963) and conformity (Asch 1955) (Bortolotti, Mameli, and Mameli 2006). Professional organisations such as the American Psychological Association, while discouraging its use, do permit deception under certain conditions. Such conditions include debriefing participants regarding the deception and allowing for withdrawal from the study after being informed of the use of deception ("Ethical Principles of Psychologists and Code of Conduct" 2016).

While there are heated debates about the use of deception, there is limited experimental evidence on its effects in the era of improved experimental methods, with some exceptions (e.g. (Barrera and Simpson 2012; Jamison, Karlan, and Schechter 2008). This is despite calls for more research (Hertwig and Ortmann 2008). In earlier experimental work on the role of false purpose deception specifically, Gallo et al (1973) find, in the context of conformity, no effect from the variations in stated experimental purpose, even when around over half of the participants in the false purpose condition were suspicious of deception.

The objective of this research is to understand the effects of varying experimental disclosures on behaviour on two widely used honesty tasks (Gerlach, Teodorescu, and Hertwig 2019; Abeler, Nosenzo, and Raymond 2019). We explore the effects of different types of experimenter disclosures, including deceptive disclosures, on a popular crowd-sourced labour platform, MTurk, where large volumes of experiments are performed and where deception is not forbidden (Mason and Suri 2012).

Given that dishonesty is generally considered to be an undesirable behaviour (Aquino and Americus 2002), it would be expected that awareness of being observed for this behaviour, as per the transparency condition, would result in lesser dishonesty (Abeler, Nosenzo, and Raymond 2019; Gneezy, Kajackaite, and Sobel 2018). Suspicion of being observed, which could be provoked by false purpose deception, may also mitigate dishonesty. Certainly, Gerlach et al (2019) found in their meta-analysis of 565 experiments that dishonesty is reduced when an (unspecified) form(s) of deception is used. However, this effect was driven by effects found in sender-receiver games and did not hold among studies using coin-flipping, die roll and matrix tasks. Finally, one could also imagine that a participant could retaliate against an experimenter by engaging in elevated cheating, if suspicious of and aggrieved by a suspected deception (Orne 1962).

A secondary objective of this work is to support those in the field who are constrained in their ability to use deception, yet unsure of the impact of ‘incomplete disclosure’ on the behaviour of interest - in my case, honesty behaviour. When undertaking replications in the field (Project 1), the relevant ethics committee and field partners would not permit me to use false purpose deception in describing experimental purpose, as per the original study (“Life and Satisfaction”). Consequently, I resorted to ‘incomplete disclosure’ (“Norms and Attitudes of Professionals”), when introducing the study. More broadly, as honesty is a commonly measured behaviour in the psychological sciences and experimental economics, and as false purpose is the most common use of deception in psychological studies, it is worthy to experimentally investigate how false purpose deception affects honesty behaviour.

A final objective is to understand what mechanisms (e.g suspicion, (in)attentiveness) may explain differences or the absences thereof on honesty which emerge from experimentally varying the stated experimental purpose. Recent research (Krasnow, Howard, and Eisenbruch 2020) found no effect from suspicion on four common experimental tasks across MTurk, economics and psychology lab samples, though older work has found some association. For example, Stricker et al. (1967) and Glinski (1970) find a relationship between suspicion of deception and reduced conformity. More broadly, others have speculation that suspicion of experimenter deception could prompt a response resistant approach from participants (Orne 1962) - in which case honesty would decline, or could motivate less socially undesirable behaviour (Gerlach, Teodorescu, and Hertwig 2019). One could also speculate that participants on MTurk, who have a monetary incentive to complete the most studies as quickly as possible, may be less attentive to disclosed experimental purpose and therefore it has minimal influence on behaviour. Alternatively, one could imagine that even if MTurkers were attending to the stated experimental purpose, even if they were suspicious of deception, that nature of the breach of trust that the deception evoked was not sufficiently egregious to prompt a change in behaviour (Smith 1981; Epstein, Suedfeld, and Silverstein 1973) or that participants may still strive to behave as if they had not been deceived, in keeping with being a 'good' and 'faithful' subject (Orne 1962; Fillenbaum 1966; Spinner, Adair, and Barnes 1977).

Chapter 2: Methods

Summary

All 10 experiments were conducted using a between-subjects design, executed via on-line surveys (using the software Qualtrics). The samples were drawn from four individual institutions and one panel (Project 1) and the crowd-sourced labour platform, Amazon Mechanical Turk (MTurk, Projects 2 and 3). Honesty was measured using tasks for whose outcomes are randomly generated - die rolls and coin flips. Participants had the opportunity to mis-report outcomes in order to increase their personal payoff or a contribution to charity. In each of the projects, assessments were made of the prevalence of honesty by comparing the reported outcomes in each condition to that predicted by the theoretical distribution of a fair coin or die roll. Analyses were made using nonparametric tests - due to skewness in the distributions - of treatment effects. Linear regressions were conducted to test for treatment effects amid other experimental and demographic controls. For Project 1, given the nature of the research, we conducted power analyses, using bootstrapping techniques, with both our and the original studies to assess the likelihood of finding treatment effects at conventional levels of significance and of replicating a treatment effect of the same or larger size as found in the original study.

Project 1

Efforts to undertake an exact replication in the original jurisdiction were frustrated by legal requirements which prevented the authors of the original study from disclosing the country where their experiments took place. Having exhausted attempts to recruit in the speculated jurisdiction, I eventually explored the replicability and generalisability of the original finding by recruiting banking and non-banking samples in what I assumed were different jurisdictions. In all, five samples from three regions were recruited - four from the field and one panel. Some samples for this project – Middle Eastern bankers (n=148) and financial services regulators (n=67) - were obtained and analysed for my Masters' thesis. Additional samples – Asian pacific bankers (n=1,178) and non-bankers (n=242 (via a panel provided by Qualtrics), European financial services regulators (n=205) - were recruited subsequently.

In the large Asia Pacific bank sample, we extended the original experimental design to add a dimension which varied the nature of the reward - either selfish or for others. That is, we also randomised whether participants, if winning the prize lottery, would receive a personal shopping voucher (self reward) or would have a donation made on their behalf to a charity associated with the bank. There was no reputational benefit from winning a charitable reward, as participants were not informed if they had ‘won’ the reward for charity and donations were not made in individual participants’ names.

Project 2

In Project 2, three studies were run. A pilot study (n=180) which enabled an assessment of whether an implicit (Gino et al. 2011) or explicit measure of self-perceived morality, adapted from Effron et al (2015) was more effective. Further, we explored whether the placement of the question of self-perceived morality in the survey affected honesty and pro-social decisions and used the honesty task findings to conduct power analyses which then guided the sample size for the main study. Briefly, we found the explicit measure of self-perceived morality to be a more sensitive measure to cheating, and determined that the placement of the self-perceived morality question did not affect the charitable donation decision. Finally, power analysis regarding correlations between the explicit self-perceived morality measure and mind-game outcomes, suggested that for a conventional power at a 5% significance level, we would need 460 participants per condition. We rounded this up to 500 for the main study.

In the main study (n=2015), participants were randomly allocated to one of four different reward conditions (max, $0.10, $0.50, $5, $50). The task is a 10-round ‘mind-game’ (Kajackaite and Gneezy 2017; Cohn, Fehr, and Marechal 2014) coin flip task, in which it is not possible to verify whether an individual cheated in any given round. The outcome of the 10-round task was paid out via a lottery mechanism from Cohn et al (2014). After completing the tasks, participants were asked to report their moral feelings, and then choose whether and how much to donate to charity. Measures were then taken of their ‘guilt proneness’, based on the GASP scale (Cohen et al. 2011) and demographics. The following day, all participants were invited back to complete a brief survey, in which we measured their self-perceived morality. The majority of participants returned (70%) to complete this study (n=1413).

Project 3

In Study 1, we randomly allocated participants to conditions which vary the nature of stated experimental purpose. Specifically, participants were randomly allocated to three conditions, with true (“Honesty”), incomplete (“Norms and Attitudes”), and deceptive (“Life and Satisfaction”) disclosures regarding experimental purpose on the welcome page and consent form. Participants then engaged in a 10-round incentivised coin-flipping task (Cohn, Fehr, and Marechal 2014). In addition to a measure of whether participants were suspicious and what they were suspicious of (free text), measures were made of experience with the honesty task and on MTurk generally, in addition to perceived past deception in other experiments.

In Study 2, we test the generalisability of the result from Study 1 with another commonly used honesty task - a die roll (Fischbacher and Föllmi-Heusi 2013). We also wanted to deepen our understanding of how suspicion of deception may affect task behaviour, so introduced a new fourth condition; an absurd false purpose condition in which we stated we were studying "Juggling Clowns." This enables us to explore suspicion as a possible mechanism to explain differences. We also introduced measures of attention (time spent on the consent form and de-briefing page) and an incentivised manipulation check which rewarded a correctly recalled stated experimental purpose. This enabled us to assess whether inattention to experimental purpose and other aspects of completing the survey could help to explain our results.

Additionally, in Study 2, we included a survey to provide insights on participant views on the use of deception. In particular, we measured their level of concern regarding differing types of deception (Sieber, Iannuzzo, and Rodriguez 1995) - the definitions of which we updated to reflect contemporary research practices. Incentivised measures were taken regarding MTurk worker peer expectations regarding the acceptability of researchers using false purpose deception on MTurk - which also gave an insight into how egregious false purpose deception was compared with other forms of deception. We also measured, depending on whether participants reported past exposure to deception or not, experienced or anticipated spillovers from exposure to deception on a range of participant behaviours (e.g. in future similar or different tasks) and attitudes (e.g. trust in science, willingness to participate in future experiments).

Chapter 3: General Discussion

Summary of findings

Project 1 Results

In contrast to the original study, we do not find treatment effects from priming banker identity in both samples. However, similar to the original study, we find in the three non-banking samples, two of which are drawn from financial services regulatory bodies - a null effect from evoking professional identity. Together, the results point to low-level variation in honesty - in both directions - that can emerge from priming professional identity.

We undertook a range of analyses in attempting to understand why the original study did not replicate. Under-powered studies are one of the causes of the reproducibility crisis (Ioannidis 2005; Munafò et al. 2017; Lane and Dunlap 1978). As such, we undertook power analyses of the original studies both with regard to finding a significant effect using the original tests and to find an effect size equal to or greater to the original finding. We found indications of inadequate power. Using bootstrapped sampling techniques, our simulations indicated that a sample of more than 170 participants is needed to achieve conventional power (80%). This compares to n=128 in the original study, and n=148, and n=620 in our banker samples. Further, the likelihood of replicating the same or larger effect size from the original study (which itself found a small effect size - *Cohen's d* = 0.37) in our banker samples was between 0.01% and 5.36%. These analyses could be used to argue that the original study was under-powered, and as such, any effect found would be unlikely to replicate beyond that specific sample and point in time. In noting indications of being underpowered, we would acknowledge the difficulties of both recruiting institutions to participate in such studies and ensuring high participation rates within institutions.

We also undertook analyses of other sampling and methodology issues that may have explained the variation between the original study's results and our replication attempts - analyses that pre-suppose a true effect from priming professional identity. Regarding sampling, one insurmountable issue with unknown magnitude is the self-selection of banks into this study after the unfavourable and highly publicised results of the original study. I approached 27 financial institutions from different regions around the world (including 14 investment banks), and only two (commercial) banks accepted to participate. In light of the reputational risks of participation, it is likely that the banks that choose to participate had a robust corporate culture.

In other sampling issues, there are likely to be differences in national culture, investment (original study) versus commercial banking (our studies) culture with respective differences in incentives, banking norms across jurisdictions, and heterogeneity in banking culture in any given jurisdiction. We find modest evidence of national culture playing a role in our smaller banking sample though not our larger banking sample, relative to the presumed jurisdiction of the original study. We find some evidence of banker honesty norms being weaker in the original jurisdiction as expectations of banker dishonesty were aligned with those of prison inmates, and higher than those of medical doctors. By contrast our sample drawn from the Asia Pacific (which seemingly has a similar national culture to the presumed jurisdiction of the original study) reveals no differences in the expectations of banker honesty relative to prisoners and medical doctors, nor the general population.[1] Further, the control group in the original study did not exhibit signs of dishonesty (unlike the control groups in our studies), suggesting that banking culture could be particularly corrosive in the original bank sampled. We cannot be conclusive in which particular aspects from various cultural and industry considerations contribute to heterogeneity in results from inducing professional identity though it is reasonable to believe that all such factors could play a role.

We also explored how a range of methodological differences could contribute to variation in results between the original and our studies. One such difference was the size of the reward. Due to differences in denominations in currencies across countries, our banker and non-banker samples had rewards of ~USD14 versus USD20 per each of the 10 rounds of the coin-flipping task. Existing literature (Gino, Ayal, and Ariely 2013; Abeler, Nosenzo, and Raymond 2019) and our own experimentation at unconventionally high levels of rewards (Project 2 - (Rahwan et al. 2018)), finds an insensitivity of honesty to large rewards.

Another methodological variation related to stated experimental purpose. The stated purpose of the original study was "Life and Satisfaction" - a form of 'false purpose' deception (Sieber, Iannuzzo, and Rodriguez 1995). Due to constraints placed on us by the overseeing ethics committee and preferences of our field partners, we used 'incomplete disclosure' when stating our experimental purpose ("Norms and Attitudes of Professionals"). Using a highly powered study, we find (as discussed further in Project 3), that there are no differences in honesty when using false purpose deception versus incomplete disclosure in stating the purpose of the experiment.

Another methodological difference introduced by field constraints related to the nature of the reward among our non-banking field samples. That is, both participating financial services regulators were

[1] I was unable to access a sufficiently sized sample to measure expectations in our smaller sample originating from the Middle East.

unwilling to participate in a study which enabled their staff to engage in unethical behaviour that would lead to them being personally enriched. Consequently, we introduced ‘charitable’ rewards wherein personal winnings would be paid on their behalf by the researchers to each organisations’ nominated charity. However, in our experimental measurement of whether a qualitative difference in rewards affected behaviour, we found no effects on honesty in the large sample of bankers. The apparent indifference between selfish and charitable rewards provides a challenge to the categorisation aspect self-maintenance theory. This aspect proposes that the ability to construe unethical behaviour in a way that avoids a negative evaluation of one’s morality will affect the propensity to behave dishonestly. In the case of charitable (versus selfish) rewards, it is arguable that there is a relative ease of recategorizing dishonest behaviour as virtuous when the outcome of that behaviour benefits others, particularly, those in need. Yet, we find no such indication of this with the coin-flipping task with high levels of stakes in this specialist population.

Project 2 Results

We find, similar to many other studies, low levels of dishonesty in each condition (Abeler, Nosenzo, and Raymond 2019). Regarding differences in honesty caused by variations in stake sizes, we find that as incentives increase, there is a small, yet significant rise in cheating. In particular, reported winning outcomes increase by less than 5% as the reward increases 500-fold, consistent with much (Fischbacher and Föllmi-Heusi 2013; Mazar, Amir, and Ariely 2008; Gneezy, Kajackaite, and Sobel 2018) literature which incorporates an aspect of internal or identity costs to engaging in dishonesty. While insensitivity of dishonesty to rewards is commonly found, it is not exclusively the case when observability is reduced to the greatest degree (Kajackaite and Gneezy 2017).

Next, we find that the propensity to donate does not vary with condition, unlike the proportion of winnings donated, which declines as the reward size increases. Only in the high stakes condition do we find a negative interaction with self-perceived morality. That is, those that feel more moral after the honesty task in the high stakes condition are less likely to donate to charity. This fits with the “moral credentials" explanation of moral licensing (Miller and Effron 2010). Having established that one is a ‘moral’ decision-maker, by refraining from claiming a large reward in the honesty task, this can subsequently be used to re-construe their subsequent ‘transgression’ of not donating a large amount to charity as ambiguous and not necessarily immoral.

We also explore how self-perceptions of morality vary immediately after the task and a day later. We find that average self-reported morality after the task does not vary by reward condition, even when interacted with the behaviour in the honesty task. Still, higher winning reports do correlate with reduced self-reported morality, consistent with our pilot study. When examining the whole sample across conditions, participants felt less moral on the day after. This appears to be driven by a group of maximal cheaters (n=169 or 8% of the sample). This group, while they thought they would be less prone to guilt from misdeeds, were the only group (of four created, based on level of reported wins) to report significantly lower feelings of morality a day later.

Project 3 Results

In Study 1, we found indications of dishonesty in all conditions, though only marginal differences between conditions. Incomplete disclosure had the highest level of mean reported coin-flip wins, followed by false purpose and true purpose. Using a non-parametric test, the difference between the two most extreme conditions was significant at conventional levels, though being less than half a coin-toss, of no economic consequence in this experimental design. Moreover, using a probabilistic model, no differences in honesty were identified. Low levels of suspicion of being deceived in general were reported (16-18%) and this could be an underestimate (Taylor and Shepperd 1996). Looking in more detail as to the nature of perceived deception, only ~2-4% of participants per condition reported a specific suspicion of false purpose.

In Study 2 we again find no effect on honesty across all four stated experimental purposes, including the absurd deception condition of “Juggling Clowns.” In the absurd deception condition, we do find elevated suspicion, however this does not translate into changed behaviour in the die roll task. We also find that the lack of variation does not appear to be driven by participant inattention to the stated purpose, with ~70% of participants across conditions correctly identifying their stated experimental purpose. Moreover, we find that while participants who were previously exposed to deception didn’t vary their behaviour in the honesty task, they spent nearly double the time on the de-briefing page and they were 13% more likely than deception-naive participants to correctly identify the stated experimental purpose. This suggests deception may increase participant attention in studies - which surprisingly is against even their own expectations. Specifically, 72% [95% CI: 69% - 75%] of participants who stated they had been previously deceived, expected no change in the level of attention paid in subsequent studies. This compared 66% [95% CI: 60% - 72%] for those reporting not past experience of deception.

While our main findings suggest that the use of false purpose deception may not affect the two experimental measures of honesty we tested, that is not to suggest that all participants are comfortable with the practice. While 78% stated that they believed false purpose deception should be permitted on MTurk, this fell to 54% when asked, using a measure incentivized for accuracy, what they thought their peers would believe. This could be interpreted to suggest, in keeping with APA regulations, that deception should be reserved as a method of last resort.

Synthesis

This thesis investigates the generalizability of the Cohn et al (2014) finding (Project 1). This highly-cited research (Altmetric – Business Culture and Dishonesty in the Banking Industry, 2020) concluded that invoking the professional identity of bankers undermined their honesty, as distinct from non-bankers. In exploring the replicability of this work we explore what factors may explain heterogeneity in effects. In support of these replication attempts, I also investigated how variation in the size of rewards – often necessitated in the field – affects honesty (Project 2). This work contributes to the existing literature by exploring, at a higher level of stakes, whether the common (Abeler, Nosenzo, and Raymond 2019), though not exclusive (Kajackaite and Gneezy 2017) finding that dishonesty is largely insensitive to rewards in experimental studies. This project also examines, through the lens of moral licensing (Miller and Effron 2010), the consequences of refraining from dishonesty amid large rewards. Finally, due to differences in university ethics requirements and field partner preferences, I could not use false purpose deception as per the original study when undertaking replication attempts. Consequently, I investigated how the nature of stated experimental purpose affects the measurement of honesty in two commonly used experimental tasks and the possible mechanisms (Project 3). This provides the first empirical insight since the 1970s regarding how the stated experimental purpose can affect a behaviour of interest.

Together, this dissertation provides novel empirical insights regarding how various contextual factors, including methodological choices by the experimenter, affect measured dishonesty, both in the laboratory and field. Broadly, I showed that factors such as making professional identity salient, changing the very nature of a reward to be selfish or charitable, varying by a very large extent rewards and the experimenter's choice of disclosed experimental purpose all have minimal impacts on the measurement of dishonesty in coin flip and die roll tasks. These findings are consistent with the notion that there is an internal cost to dishonesty - in terms of one's self-conception as an ethical actor and possibly the perception from others of the same - and that this dominates both conscious considerations in the decision to behave dishonestly such as reward size, the nature of the reward, knowledge of what the experimenter is measuring, and unconscious influences such as priming professional identity.

Replicability of dishonesty research

The ability to replicate results is key to building trust in science, progressing scientific understanding and supporting policy-making. Similarly, replication attempts - both direct and conceptual - are important to understand the limits to generalising findings and identifying factors that drive heterogeneity in effects, especially when they are novel and hold a significant profile in academic and public circles. I hope that our studies with bankers, regulators and non-bankers help to further the understanding of how professional identity may influence honesty (Project 1). Specifically, we suggest that, given one assumes there is a true effect from priming professional identity, that there is heterogeneity in banking and other professional cultures, and that this can cause variation in measured honesty. While it is difficult to identify the precise cause of variation in honesty from inducing banker professional identity, especially given the small effect size from the original study and our negligible effect sizes, our strongest evidence points to variation in national banking norms. Moreover, I hope that our work on reward sizes (Project 2) and experimental methodology (Project 3) assist other researchers in managing constraints they may face in the field as they strive to undertake faithful replications.

Our attempts to faithfully replicate earlier research was challenged by restrictions on the use of false purpose deception while measuring honesty in the field. Earlier conceptual work following controversial experiments in the period during and soon after World War II and a recent meta-analysis of honesty studies suggests that the presence of deception has consequences for participant behaviour. We do not find experimental evidence of this (Project 3), at least of how false purpose deception affects our measures of honesty on MTurk. However, there remains a significant gap in understanding how a range of behaviours in tasks are affected by various types of deception, and this may help to understand heterogeneity of various effects more broadly. Further, better documentation of participants' past exposure to deception - which we understand to be frequently deficient in both on-line experimental platforms and physical laboratories - may also assist researchers better understand the nature of their experimental samples, including possible biases emerging from selective attrition.

The role of internal costs in dishonesty

Traditional models of dishonesty (e.g. Becker, 1968), which take no account of the internal costs of unethical conduct, would predict that increased sizes of rewards from dishonesty would induce more dishonesty. Moreover, they would predict greater dishonesty when the reward was for oneself versus for charity. However, in keeping with more recent models which incorporate such costs (e.g. Mazar et al., 2008, Gneezy et al., 2018), we show that variation in the size of a reward (Project 2) and the nature of the reward (Project 1) have a negligible effect on dishonesty. With regard to the charitable reward, one can conceive that is an external reward of $0 while the internal reward would also be equivalent to zero given the donation was not made in winners' names and winners were not informed of any donation made, so they would be unable to signal their charitable contribution to peers or use it to maintain their self-concept as a moral person. The former study helps to generalise existing findings at high levels of rewards and we believe that the latter study provides a novel contribution to the literature, and from a unique population. We hope that together these findings are both of theoretical interest - and of practical benefit for researchers in the field, who like us, could not always precisely match the value of rewards (e.g. due to different denominations of currencies), or are unable to offer selfish rewards to particular populations.

A question remains as to the nature of internal costs. In particular, do internal costs vary as the size of the lie increases? Our direct measures of moral self-perceptions (Project 2) suggest that there is no variation in internal costs, on average, as the size of rewards vary 500-fold. Self-concept maintenance theory Mazar et al. (2008) could be interpreted to suggest that at high stakes, dishonesty is harder to engage in without triggering a negative re-appraisal of oneself. Despite this, we find no indication of this based on the level of rewards available nor the nature of rewards (Project 1). However, examining individual variation across the sample in Project 2, independent of the size of reward, we find support for this notion. Specifically, the subset of those engaged in maximal cheating while not feeling worse on the day of their maximally unethical behaviour, they were the only group to feel more negative moral self-perceptions a day later - something which they mis-predicted. This highlights some interesting individual variation and temporal aspects to internal costs which is not to my understanding reflected in existing models (e.g. (Gneezy, Kajackaite, and Sobel 2018).

The consequence of refraining from dishonesty amid large rewards

In Project 2, we also showed that there are some specific spillovers on pro-social behaviour from refraining from dishonesty. Specifically, the proportion of winnings donated to charity declined as the scale of winnings increased, and only in the condition with the highest rewards was there evidence that individuals leveraged their moral feelings to engage in less charitable behaviour. This is in line with what would be predicted by psychological licensing literature (Miller and Effron 2010; Zhong, Liljenquist, and Cain 2009), and specifically, the "moral credentials" pathway. More broadly, while we have shown low levels of cheating (Projects 1, 2, 3) - consistent with the existing literature - there may be adverse spillovers that are being missed by standard experimental paradigms in which spillovers on subsequent behaviours and feelings are not commonly measured. The finding of adverse spillovers of subsequent pro-social behaviour also has important practical implications for the design of organisations. For example, these findings suggest that to avoid adverse spillovers from resisting temptations that would personally enrich staff, it is preferable to eliminate the temptation, where practicable, rather than to engage in deterrence from being tempted. Further, one could conceive of amending decision making processes in a manner which spaces decisions for high stakes decisions (e.g. those in senior management or at Board level) to minimise the chance of adverse spillovers on subsequent decisions.

Caveats and Considerations

Broadly, our work in the field (Project 1) highlights the difficulties in conducting high fidelity replications of well-publicised results. In particular, the self-selection of more 'ethical' banks into our research also provides an insurmountable threat to replication. Despite intensive efforts over a number of years, we were only able to recruit two banks of 27 in which we held detailed discussions regarding participation. This selection problem could only be overcome by testing for generalisability of an original finding ahead of publication, and potentially refraining from sharing existing results with prospective field partners - which itself poses an ethical dilemma. Of course, testing generalisability of findings is not always feasible, especially in studies with difficult-to-access populations and on sensitive topics like honesty which could pose reputational risks to the organisation. Nevertheless, I would advocate for such a process where practicable.

Beyond the difficulties in recruiting institutions, it is important to note that there are a range of other factors that cannot necessarily be controlled in the recruitment of participants within institutions. For example, the within-institution non-participation rates varied notably across institutions (26%-56%) which we expect may have been driven by variation in the manner in which the survey was distributed. The variation in sample could emerge from who initiated the invitation (e.g. CEO, Communications team) and the method of distribution (e.g. all staff email, intranet posting). Variation may also emerge from the disclosures required by different academic institutions in participant consent forms. While the availability of experimental materials, data, code, etc has greatly improved, for field work in the social sciences there is an opportunity to encourage better practices which may ultimately capture systemic differences generated by factors beyond experimenters' control.

Reflecting on the studies involving crowd-sourced labour platform samples (Projects 2 and 3), our findings could have been strengthened by undertaking representative sampling on MTurk regarding standard demographic measures (age, gender, ethnicity). On MTurk, using an intermediary like Prime Panels, a representative sample can be obtained, at a higher cost. Still, the overall data quality from such panels may be inferior, in the absence of screening or attention checks. Moreover, there is variation in compensation to participants as determined by sample providers (Chandler et al. 2019). Given the noted variation in participant behaviour on popular platforms such as MTurk, CrowdFlower and Prolific (Peer et al. 2017; Gupta, Rigotti, and Wilson 2021) and emphasis on honesty in on-boarding Prolific participants, our results may not be representative for all on-line platforms, let alone the general population. Finally, our findings may not hold over time. While this issue was minimised in Project 3 given the multi-year gap between studies, the estimated 50% turnover of the MTurk population every seven months (Stewart et al. 2015) raises questions over the stability of results on this platform.

There are also considerations regarding the nature of tasks used to measure honesty. For the replication studies we used a coin-flip task which tends to reveal lower rates of dishonesty than other common honesty tasks - die rolls, sender-receiver games and matrix tasks according to meta-analysis (Gerlach, Teodorescu, and Hertwig 2019) though not controlled experimental study has been undertaken comparing performance with the same sample(s) as far as I am aware. Further, the tasks we used have been classified as ‘self-reported outcomes’ and the results from other classes of tasks (e.g. lost items, undeserved money) (Rosenbaum, Billinger, and Stieglitz 2014) may result in differing conclusions - via interaction with different treatments. As such, findings from our coin flip and die roll tasks used in all projects may not generalise to other types of honesty tasks.

Future Directions

The continuation of scandals in the banking industry and the large consequences for society, ensures that the role of professional culture in driving dishonesty remains a research topic of importance. My work points to the importance of on-going efforts to bring experimentation to banking and other industries to better measure and understand aspects of professional culture, such as, but not limited to, honesty. Given the high-profile nature of the original paper on banking culture and its effect on honesty, this approach to assessing honesty is vulnerable to being gamed by participating institutions, especially if, for example, they were mandated by regulators. Rather, there is a need for developing new tools to measure dimensions of professional culture. For some types of financial services, this could be in the form of an experimental approach to ‘mystery shoppers’ or other approaches like the ‘lost wallet’ paradigm (Milgram, Mann, and Harter 1965), recently adapted to measure honesty in a large, cross-cultural field study (Cohn et al. 2019).

New tools are also needed to counter dishonesty. The exploratory analyses of maximal cheaters in Project 2 show that they were the only group to mis-predict their tendency to experience negative moral feelings after a transgression. While the effect is small, it does raise the question of whether providing an ethical reminder or boost (Hertwig and Grüne-Yanoff 2017) to those most likely to engage in a (large) transgression that they are likely to feel (more) poorly about such behaviour than they expect, may be an effective deterrent to unethical behaviour.

A number of methodological questions remain open regarding the measurement of honesty. One line of research is how generalisable findings from one honesty task are to other honesty tasks. That is, to what degree are findings dependendent on the nature of the honesty task used? While existing meta-analyses are helpful in providing some insights, there is a shortage of experimental work on this topic. This could also further the understanding of individual differences in the propensity to engage in dishonesty across different tasks.

Another promising line of research relates to the use of deception. In our studies, we explored the use of false purpose - one of eight types of deception (Sieber, Iannuzzo, and Rodriguez 1995), on two different honesty tasks. I hope our work inspires future research on how false purpose may affect other types of commonly measured behaviours (e.g. cooperation). More broadly, one could conceive of an extensive research agenda which explores experimentally how other types of deception may affect the measurement of various experimental behaviours (e.g. suspicion, willingness to participate in future studies) and attitudes (e.g. trust in science and researchers). While this would require care with any use of deception, subsequent results could bring some much needed empirical insights to on-going tensions between disciplines regarding the use of deception.

Chapter 4: Manuscript of Project 1

Rahwan, Z., Yoeli, E., & Fasolo, B. (2019). Heterogeneity in banker culture and its influence on dishonesty. Nature, 575(7782), 345-349. https://doi.org/10.1038/s41586-019-1741-y

Zusammenfassung auf Deutsch

Die Sozialwissenschaften befinden sich in einer so genannten "Reproduzierbarkeitskrise.“ Sehr einflussreiche Ergebnisse aus zugänglichen Populationen, wie Laboratorien und Crowd Sourced-Worker-Plattformen, werden nicht immer repliziert. Weniger Aufmerksamkeit wurde der Replikation von Ergebnissen aus unzugänglichen Populationen gewidmet, und in der Tat schlossen die jüngsten hochkarätigen Replikationsversuche solche Populationen ausdrücklich aus. Eine bahnbrechende experimentelle Arbeit bot einen seltenen Einblick in die Kultur von Bankern und fand heraus, dass Banker, im Gegensatz zu anderen Berufsgruppen, unehrlicher sind, wenn sie über ihren Job nachdenken. Angesichts der Bedeutung des Bankensektors ist eine Untersuchung ihrer Verallgemeinerbarkeit gerechtfertigt, bevor sich die Wissenschaft oder politische Entscheidungsträger auf diese Ergebnisse als genaue Diagnose der Bankenkultur verlassen. Hier führen wir die gleiche incentivierte Aufgabe in fünf verschiedenen Populationen, über drei Kontinente mit 1.282 Teilnehmern durch. In zwei Studien (n=148, n=620) beobachten wir eine gewisse, wenn auch nicht signifikant erhöhte Unehrlichkeit unter Bankern, die zum Nachdenken über ihre Arbeit angeregt wurden. Wir finden auch, dass es keinen signifikanten Effekt auf die Ehrlichkeit hat, wenn Nicht-Banker (n=67, n=205, n=242) über ihre Arbeit nachdenken. Wir untersuchen Stichproben- und methodische Unterschiede, um die Variation der Ergebnisse in Bezug auf Banker zu erklären und identifizieren zwei Schlüsselpunkte: die relativen Erwartungen der Allgemeinbevölkerung an das Verhalten von Bankern variieren von Land zu Land, was darauf hindeutet, dass die Bankenkultur in der ursprünglichen Jurisdiktion möglicherweise nicht länderübergreifend konsistent ist und da wir 27 Finanzinstitute angesprochen haben, von denen viele Bedenken hinsichtlich negativer Ergebnisse äußerten, erwarten wir, dass nur Banken mit einer soliden Kultur an unserer Studie teilgenommen haben. Letzteres birgt ein erhebliches Risiko von Selektionsverzerrungen, die die Verallgemeinbarkeit jeder ähnlichen Feldstudie untergraben können. Im weiteren Sinne zeigt unsere Arbeit die Komplexität der Replikation einer sensiblen, öffentlichkeitswirksamen Feldstudie, die aufgrund institutioneller und geographischer Barrieren für die Bevölkerung kaum zugänglich ist. Für politische Entscheidungsträger legt diese Arbeit nahe, dass sie bei der Verallgemeinerung der Ergebnisse auf ihren nationalen Zuständigkeitsbereich Vorsicht walten lassen sollten.

Article

Heterogeneity in banker culture and its influence on dishonesty

https://doi.org/10.1038/s41586-019-1741-y

Received: 9 September 2016

Accepted: 26 August 2019

Published online: 13 November 2019

Zoe Rahwan[1,2,3]*, Erez Yoeli[4] & Barbara Fasolo[2]

The social sciences are going through what has been described as a 'reproducibility crisis'[1,2]. Highly influential findings derived from accessible populations, such as laboratories and crowd-sourced worker platforms, are not always replicated. Less attention has been given to replicating findings that are derived from inaccessible populations, and recent high-profile replication attempts explicitly excluded such populations[3]. Pioneering experimental work[4] offered a rare glimpse into banker culture and found that bankers, in contrast to other professionals, are more dishonest when they think about their job. Given the importance of the banking sector, and before academics or policy-makers rely on these findings as an accurate diagnosis of banking culture, an exploration of their generalizability is warranted. Here we conduct the same incentivized task with bankers and non-bankers from five different populations across three continents ($n = 1{,}282$ participants). In our banker studies in the Middle East and Asia Pacific ($n = 148$ and $n = 620$, respectively), we observe some dishonesty, although—in contrast to the original study[4]—this was not significantly increased among bankers primed to think about their work compared to bankers who were not primed. We also find that inducing non-banking professionals to think about their job does not have a significant effect on honesty. We explore sampling and methodological differences to explain the variation in findings in relation to bankers and identify two key points. First, the expectations of the general population regarding banker behaviour vary across jurisdictions, suggesting that banking culture in the jurisdiction of the original study[4] may not be consistent worldwide. Second, having approached 27 financial institutions, many of which expressed concerns of adverse findings, we expect that only banks with a sound culture participated in our study. The latter introduces possible selection bias that may undermine the generalizability of any similar field study. More broadly, our study highlights the complexity of undertaking a high-fidelity replication of sensitive, highly publicized fieldwork with largely inaccessible populations resulting from institutional and geographical barriers. For policy-makers, this work suggests that caution should be exercised in generalizing the findings of the original study[4] to other populations.

A previous experimental study[4] has found evidence that when investment bankers are reminded of their professional identity (treatment), they become more dishonest than their colleagues who are asked to think about leisure activities (control). No such effect is found when priming professional identity among non-banking professionals. Together, these results were interpreted as providing evidence that banking identity was associated with weaker honesty norms, and were cited widely, in both academic literature[5–8] and the media[9–11].

In this experimental paradigm, honesty is measured via a simple coin-flip task. After answering questions about either their professional identity or leisure activities, bankers are asked to report the outcomes from 10 flips and are paid around US$20 for each reported win. Without cheating, bankers should report winning coin tosses 50% of the time, on average, and the variation in reported wins should be characterized by a binomial distribution. Of course, it is impossible to infer how much more bankers cheat in real-world large stakes, although there are positive correlations between honesty experiments and real-world outcomes[12–15].

The presence of weaker honesty norms in banking culture has considerable negative implications for society collectively, as demonstrated by the role that dishonesty played in the subprime mortgage crisis[16–18]. This concern remains current: since the original study[4] was conducted, further scandals have emerged in both investment and commercial banks[19–25], trust in banking professionals and banks remains at relatively low levels[26,27] and policy-makers remain concerned about culture in the banking industry[28,29]. In light of this concern, as well as

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. [2]Department of Management, London School of Economics and Political Science, London, UK. [3]Harvard Kennedy School, Harvard University, Cambridge, MA, USA. [4]Sloan School of Management, MIT, Cambridge, MA, USA. *e-mail: zrahwan@mpib-berlin.mpg.de

ongoing concerns regarding the reproducibility of experimental results in the social sciences[2,3,30–33] and limited representativeness regarding national cultures[34], we explore the generalizability of this influential field study of bankers.

In our first study, we used the design of the original study[4] ($n = 128$ with a follow-up study of $n = 80$) on a considerably larger sample of 620 commercial (not investment) bankers at a large bank in the Asia Pacific region. Bankers in the treatment group tended to be more likely to cheat than the control group, although the difference was smaller in the Asia Pacific sample than in the original study (Cohen's[35] $d = 0.06$ compared with 0.37 in the previous study[4]) and not statistically significant at conventional levels (54.0% among primed bankers versus 53.0% among non-primed bankers, $P = 0.111$ in the present study compared with 58.2% among primed bankers versus 51.6% among non-primed bankers, $P = 0.017$ in the original study; one-tailed Wilcoxon rank-sum tests, as are all subsequent results unless stated otherwise; Extended Data Fig. 1a and Extended Data Table 1a). Among a sample of full-time and part-time non-banking employees ($n = 242$), in which we strove to be nationally representative for gender and age, we found—as described previously[4]—that those primed with professional identity were no more likely to cheat than their non-primed counterparts (58.5% versus 54.8%, $P = 0.114$ compared with 55.8% versus 59.8%, $P = 0.936$ in the original study[4]; Extended Data Fig. 1b and Extended Data Table 2a). Notably, the direction of the Asia Pacific non-banker effect was consistent with those of the banker samples and opposite to the direction of the non-banker effect in the original study[4]. Furthermore, the effect size of treatment on Asia Pacific non-bankers (Cohen's $d = 0.19$) was larger than that for bankers within the same jurisdiction.

In our second study, we used the same study design on commercial bankers at a medium-sized bank in the Middle East ($n = 148$). Bankers in the treatment group tended to be more likely to cheat, although—similar to the Asia Pacific bankers—the effect was very small (Cohen's $d = 0.11$) and not statistically significant (56.9% versus 54.9%, $P = 0.261$; Extended Data Fig. 2a and Extended Data Table 1b). In a small sample ($n = 67$) of regulators of financial services (that is, non-bankers) in the same region, we again found no treatment effect from priming professional identity (50.3% (treatment), 51.1% (control), $P = 0.472$; Extended Data Fig. 2b). The direction of the effect was aligned with that of the non-bankers in the original study[4], although not with Asia Pacific non-bankers. In a larger sample of European non-bankers ($n = 205$), who were also regulators of financial services, we again found no significant effect after priming professional identity (52.2% (treatment), 52.6% (control), $P = 0.572$; Extended Data Fig. 3). We note that for the Middle Eastern and European non-banker studies, we were constrained to use rewards for charity rather than for the participants, although similar to previous research[36], we found in a separate study ($n = 1{,}179$) that this probably did not affect honesty (Supplementary Information 1.2 and Extended Data Figs. 4, 5).

In summary, we do not find any significant increase in dishonesty among primed bankers in the Middle East and Asia Pacific, in contrast to the main study of the original publication[4], although the results trend in the same direction. Consistent with the original study[4], we find no significant effects on dishonesty from having non-banking professionals think about their jobs. Together, the findings of the original study[4] and our studies reveal that inducing professional identity results in varying effects on honesty across professions and jurisdictions, both in direction and range ([−4.0, +6.6] percentage point difference in average winning outcomes).

We do find a detectable increase in dishonesty when bankers are reminded of their profession when pooling data from our studies with the previous main study[4] ($P = 0.018$) and when pooling with the original main and follow-up studies of the previous study ($P = 0.008$) (Fig. 1 and Extended Data Table 1c, d). The effects among all banker samples, although small, are all in the same direction. However, the findings of the original main study[4] did not replicate based on the conventional significance level of $\alpha = 0.05$ in the original follow-up study, the Middle Eastern and Asia Pacific individual banker samples or when we pooled the Middle Eastern and Asia Pacific groups ($P = 0.082$). This suggests that the findings of the original main study[4] are not generalizable beyond the original population sampled. Further, variation in the direction of priming effects among non-banker samples raises the question of how likely the differences found in the original studies between bankers and non-bankers are to replicate, at least outside of the original jurisdiction.

Although the above-reported findings suggest that there may not be a consistent, or sizeable, adverse effect of priming the professional identity of bankers on honesty, we nevertheless undertake exploratory analyses to understand how sampling and other methodological and statistical power reasons may explain why our results differ from those of the original study[4].

Regarding sampling issues, one possibility for explaining the variation in results is that the culture among bankers in our studies is different from the culture among bankers in the undisclosed location of the original studies. This could result from variations in national norms[5], banker norms between countries, and/or self-selection of bankers into differing industries (for example, investment versus commercial banking). For national norms, we find negligible differences between the presumed original jurisdiction (which cannot be disclosed for legal reasons) and our Asia Pacific jurisdiction. However, some differences between the original and Middle Eastern jurisdictions are identified, which may help to account for the increased baseline dishonesty found among non-primed Middle Eastern bankers and consequently, the smaller effect size from inducing professional identity (Supplementary Information 2.3.1).

For the banking norms, we do not find evidence of differences in honesty among treated bankers across the three jurisdictions (Supplementary Information 2.3.2), although we do identify differences in people's relative expectations of bankers—a potential indicator of the heterogeneity in national banking norms. Specifically, bankers in the Asia Pacific jurisdiction are not expected by others in the same jurisdiction to be more or less honest than doctors, prison inmates and the general population (Supplementary Information 2.2.6). This is in contrast to the jurisdiction in the original study[4] in which bankers were perceived to be less honest than doctors, tended to be less honest than the general population and were indistinguishable from prison inmates (Fig. 2 and Extended Data Fig. 6).

Heterogeneity in banking culture across segments of the industry is also reasonable to expect, given the variance in remuneration structures, business models and clients. Indeed, there are perceptions of lesser honesty among investment bankers relative to commercial bankers[37,38]. Despite this, we find no evidence that investment banking culture is more aversive than that of commercial banks (Supplementary Information 2.3.3).

Another source of variation in results that we explored was the possible self-selection of more honest people into banking rather than non-banking jobs in the original jurisdiction. Among the non-primed participants of the original study[4], non-bankers engage in greater dishonesty than bankers (59.8% compared to 51.6%, $P = 0.002$)[4], in contrast to our findings in other jurisdictions (Supplementary Information 2.3.2). Furthermore, the non-primed bankers in the original study[4] contrast with our banker samples in that they do not engage in statistically detectable dishonesty (Supplementary Information 2.1.4). And this is despite the poor expectations of bankers relative to others in that jurisdiction. The absence of statistically detectable dishonesty among untreated bankers of the original main study[4]—an indication of more honest people self-selecting into the industry—may contribute to the larger effect of priming that was identified. And this in turn suggests, in the context of other banker studies with smaller effect sizes, that banking culture in the original jurisdiction may have a more aversive effect on honesty norms than elsewhere.

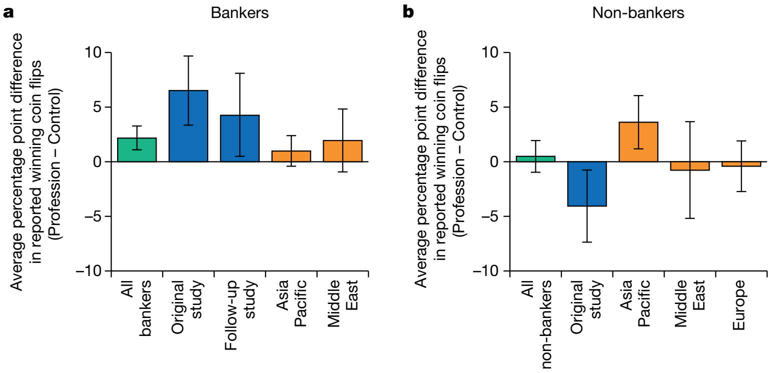


Fig. 1 | Variation in how priming professional identity affects honesty. **a**, Comparisons of the direction and size of effect from priming professional identity across separate samples of bankers around the world and the combination of all of the bankers. Although all studies find that making banker identity salient induced greater dishonesty, only the original study ($P = 0.017$, one-tailed Wilcoxon rank-sum test; $n = 128$, Cohen's $d = 0.37$) and the pooling of all bankers ($P = 0.008$, one-tailed Wilcoxon rank-sum test; $n = 976$, Cohen's $d = 0.127$) find a significant yet 'negligibly sized' treatment effect. This effect is not found in the follow-up study ($n = 80$) from the original paper ($P = 0.097$, one-tailed Wilcoxon rank-sum test; $n = 80$, Cohen's $d = 0.26$) or among the individual samples of Asia Pacific bankers ($P = 0.111$, one-tailed Wilcoxon rank-sum test; $n = 620$, Cohen's $d = 0.06$) and Middle Eastern bankers ($P = 0.261$, one-tailed Wilcoxon rank-sum test; $n = 148$, Cohen's $d = 0.11$). **b**, Comparisons of the size of effect from priming professional identity across separate samples of non-bankers around the world and the combination of all of the non-bankers. Making professional identity salient induced different directions of effects among individual samples of non-bankers from the original study, Asia Pacific, Middle East and Europe, none of which were statistically significant ($P = 0.128$ ($n = 133$), $P = 0.114$ ($n = 242$), $P = 0.472$ ($n = 67$) and $P = 0.572$ ($n = 205$), respectively; one-tailed Wilcoxon rank-sum tests). Data are mean ± s.e.m.

At an institutional level, a self-selection bias may assist in explaining the variation in results across the banker studies. The concern of adverse findings among banks appears to have biased participation towards banks that are more likely to have a sound culture. In total, we approached 27 financial institutions around the world (including 14 investment banks), and only 2 (commercial) banks accepted the invitation to participate. Highlighting the self-selection bias, one lawyer specializing in compliance with investment-banking clients who assisted us with recruitment noted regarding the difficulties that we experienced when asking institutions to join our research that "...I am particularly disappointed because the main reluctance I have encountered is clearly a concern by different firms that the survey might identify weaknesses in their culture that they are worried might somehow be exposed." One possible way around this for future studies, especially on sensitive topics, would be for the work to be replicated by the same researchers or other groups in the same or other jurisdictions, ahead of publication of the initial findings.

Publicity surrounding the original study[4] may also have affected participant responses in our study and any future studies that will follow. In the Asia Pacific banker sample, we found that 30% of respondents reported familiarity with research on banker culture that used a similar survey, although this does not appear to account for the smaller effect size relative to the original study[4] (Supplementary Information 2.1.5).

Beyond these sampling and familiarity concerns, we explored how a range of methodological differences with the original study[4] affected variation in results.

One notable methodological difference concerns the experimenter disclosure about the purpose of the experiment. Primarily driven by strict ethical rules governing our research, we used 'incomplete disclosure' in our studies rather than deception, as was used in the original study[4,39]. In a separate study, we found no statistical differences in winning outcomes in the same coin-flipping task between 'incomplete disclosure' ($n = 309$) and deception ($n = 315$) conditions (59.9% versus 58.4%, $P = 0.188$ (two-tailed Wilcoxon rank-sum test); Supplementary Information 2.4 and Extended Data Figs. 4, 5). This provides an indication that variation in experimenter disclosure may not account for differences in effect sizes, although it may not generalize to bankers in all sampled jurisdictions. Still, a meta-analysis of dishonesty experiments finds less dishonesty associated with experiments using deception, and no statistical differences with coin-flipping tasks in particular[40].

In order to have rewards in line with the denominations of local currencies, we adjusted the reward amounts from US$20 per coin toss in the original study, to approximately US$14 per coin toss in all four Asia Pacific and Middle Eastern studies. We have reason to believe that the lower rewards did not have a large effect on the variation in results. We extended the Asia Pacific study to randomize the opportunity to win a reward for oneself or for charity—effectively US$0 for the individual—and found no difference in dishonesty levels (Supplementary Information 1.2 and Extended Data Figs. 7, 8). This complements an increasing number of studies that have found that dishonesty is largely insensitive to rewards[36,41]—including extraordinarily large rewards[42].

A number of other factors that we could not control and that may account for differences in results include who within participating organizations sent out the invitations to complete the survey, how the survey was circulated and the contents of the consent form. Responses could be influenced by whether an invitation comes from a CEO (chief executive officer)—as was the case in the Asia Pacific sample—or from a member of a communications team—as in our Middle Eastern sample—or via an alumni network, as was the case in the previously published follow-up study[4]. In addition, sampling can be affected by the nature of the invitation, such as all staff being emailed or a posting on an internal corporate website. Furthermore, the wording of the consent form probably differed as academic institutions have varying requirements. The effects of these differences on all of the banker responses are indeterminable and highlight the difficulty of conducting tightly controlled replications of studies in the field.

Finally, the timing of the studies is another factor that may affect outcomes. Since the 2008 financial crisis, global standard setters such as the Basel Committee for Banking Supervision have undertaken considerable efforts, which cascade to national regulators, to deter untoward institutional and individual behaviour in the banking industry that can lead to financial instability and unfair outcomes. Assuming that these efforts have been globally efficacious, one would expect a diminished effect size on banker populations studied subsequent to the previous study[4].

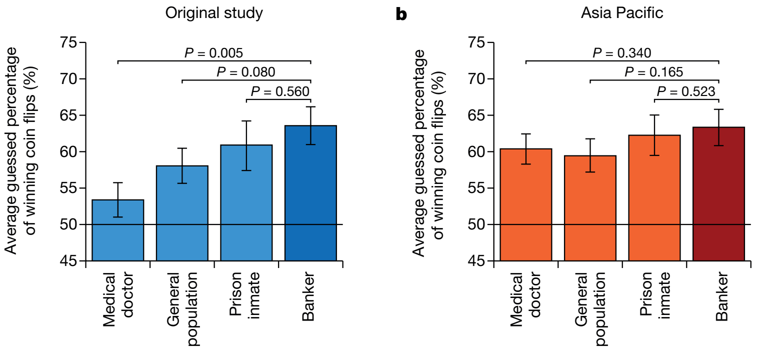


Fig. 2 | Relative expectations of banker behaviour within and across jurisdictions. a, The original study ($n = 183$) provided evidence that expectations of banker behaviour ($n = 48$) in the coin-flipping task are indistinguishable ($P = 0.558$, two-tailed Wilcoxon rank-sum test) from prison inmates ($n = 45$). Furthermore, bankers were perceived to be less honest than doctors ($n = 44$, $P = 0.005$, two-tailed Wilcoxon rank-sum test) and tended to be perceived as less honest than the general population ($n = 46$, $P = 0.080$, two-tailed Wilcoxon rank-sum test). Expectations were sourced in a convenience sample of visitors to a Municipal Office. Only men were included in the sample reported. **b**, In the Asia Pacific jurisdiction, using the same two-tailed Wilcoxon rank-sum tests, we found no statistical differences between expectations of banker behaviour ($n = 65$) and other groups assessed in the original study—prison inmates ($n = 64$), the general population ($n = 58$) and medical doctors ($n = 55$). This suggests that relative expectations of bankers vary between jurisdictions. Differences in sampling may also account for the variation in findings. In contrast to the original study[4], we sourced a nationally representative panel, by age and gender. Data are mean ± s.e.m.

Statistical power is another factor that may account for the variations among the different studies. In the social sciences, the 'replication crisis' has in part been driven by issues of inadequately sized samples[43–47]. This problem is exacerbated when conducting experiments in the field, where there is often a limited ability to access both institutions and individuals. Indeed, some replication efforts deliberately averted studies that involve 'inaccessible' populations, such as bankers[3]. Here, we find that it is difficult to find a significant effect without accessing larger samples (Fig. 3a). Specifically, a sample of more than 170 individuals is required to achieve conventional power of 80% (Supplementary Information 3.1).

One established consequence of under-powered studies is smaller effect sizes in subsequent replications[3,32,33,44,45]. The original main study[4] found a 'small' (Cohen's $d = 0.37$) effect of banker priming with an average of approximately 0.7 more winning coin flips reported for such bankers. Consistent with replication trends, we find smaller effects from priming in our studies; on average 0.1 and 0.2 winning coin flips reported by primed bankers in our Asia Pacific and Middle Eastern studies, respectively (Cohen's $d = 0.06$ and 0.11). Furthermore, bootstrapped simulations using samples of varying sizes confirm a low likelihood of finding an effect size equal to or greater than the original study (Fig. 3b and Supplementary Information 3.1).

Our studies highlight a number of acute challenges in replicating field studies at high fidelity with inaccessible populations on the question of whether banker culture undermines honesty. These new results suggest that the original finding, which led some to conclude that banking culture is 'corrosive'[48], does not appear to generalize across countries, banking segments, individual institutions or time. Although we find a negligible influence of banker identity on dishonesty, in contrast to the original study[4], it is difficult to precisely identify the underlying causes. Our strongest evidence points to differences in national banking norms and a pronounced threat of only 'ethical' banks who agreed to

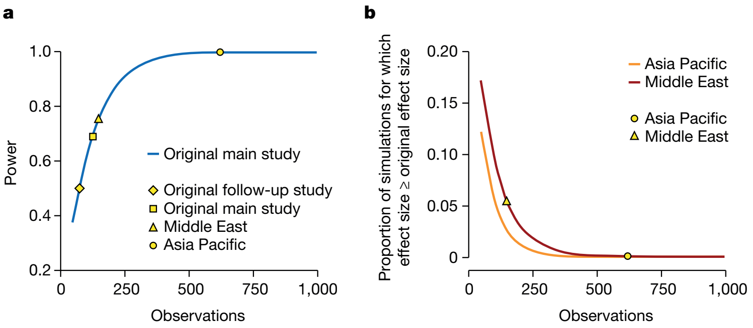


Fig. 3 | Replicability of the original study. a, The original main study ($n = 128$) has an estimated power of approximately 68% on the basis of bootstrapped samples subjected to one-tailed Wilcoxon rank-sum tests. That is, of 10,000 simulated samples, around 68% were found to have a statistically significant result given $\alpha = 0.05$. This is below the 80% level of power that is conventionally targeted—a level that would require approximately 170 participants. The chart also marks the sample sizes of the original follow-up study ($n = 80$), Asia Pacific ($n = 620$) and Middle Eastern ($n = 148$) bankers on the power curve. These samples would have a power of around 51%, 99% and 75%, respectively, on the basis of bootstrapped samples from the original main study. This suggests that the Asia Pacific sample is adequately powered to detect a treatment effect at conventional levels, although the Middle East sample is not. **b**, The original study found a difference of 0.7 average coin flips between the control and treatment groups. Drawing simulated samples of various sizes from Asia Pacific and Middle Eastern bankers reveals a low likelihood of finding the same or larger effect size than the original study in each of those samples (0.01% and 5.36%, respectively).

participate in such research—a threat that cannot be effectively countered after high-profile press coverage of the original research. Finally, we must highlight a plausible parsimonious explanation: the effect observed in the original study[4] may only have held in a very specific setting and point in time and, as such, does not generalize beyond the sample contained in their main study, nor across time.

Irrespective of the precise sources of the differences observed in the various studies of banker honesty, our findings have broad implications for replicability and generalizability outside of commonly accessed experimental populations. While, theoretically, such variation could be better understood with large-scale testing within and across countries coordinated by banking regulators, such an approach would invite gaming from institutions in an effort to protect their reputations. Instead of focusing on direct replications, we believe that new tools and methodologies are needed to measure aspects of professional culture, such as honesty. This will be critical to better understand and ultimately manage the related risks and benefits to society that stem from the banking industry.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1741-y.

1. Pashler, H. & Wagenmakers, E. J. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* **7**, 528–530 (2012).
2. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
3. Camerer, C. F. et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav*. **2**, 637–644 (2018).
4. Cohn, A., Fehr, E. & Maréchal, M. A. Business culture and dishonesty in the banking industry. *Nature* **516**, 86–89 (2014).
5. Gächter, S. & Schulz, J. F. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531**, 496–499 (2016).
6. Purzycki, B. G. et al. Moralistic gods, supernatural punishment and the expansion of human sociality. *Nature* **530**, 327–330 (2016).
7. Zingales, L. Presidential address: does finance benefit society? *J. Finance* **70**, 1327–1363 (2015).
8. Benjamin, D. J., Choi, J. J. & Fisher, G. W. *Religious Identity and Economic Behavior.* Report No. w15925 (National Bureau of Economic Research, 2010).
9. Lying, cheating bankers. *The Economist* (22 November 2014).
10. Cookson, C. Bankers have tendency to lie for financial gain, say scientists. *Financial Times* (19 November 2014).
11. Goodley, S. Coining it in: banking industry culture promotes dishonesty, research finds. *The Guardian* (19 November 2014).
12. Halevy, R., Shalvi, S. & Verschuere, B. Being honest about dishonesty: correlating self-reports and actual lying. *Hum. Commun. Res*. **40**, 54–72 (2014).
13. Dai, Z., Galeotti, F. & Villeval, M. C. Cheating in the lab predicts fraud in the field: an experiment in public transportation. *Manage. Sci*. **64**, 1081–1100 (2018).
14. Kröll, M. & Rustagi, D. *Shades of Dishonesty and Cheating in Informal Milk Markets in India*. SAFE Working Paper Series No. 134 (EconStor, 2016).
15. Potters, J. & Stoop, J. Do cheaters in the lab also cheat in the field? *Eur. Econ. Rev*. **87**, 26–33 (2016).
16. The United States Department of Justice. *Bank of America to Pay $16.65 Billion in Historic Justice Department Settlement for Financial Fraud Leading up to and During the Financial Crisis* https://www.justice.gov/opa/pr/bank-america-pay-1665-billion-historic-justice-department-settlement-financial-fraud-leading (2014).
17. SEC. *SEC Enforcement Actions Addressing Misconduct That Led To or Arose from the Financial Crisis* https://www.sec.gov/spotlight/enf-actions-fc.shtml (Securities and Exchange Commission, 2016).
18. United States Senate Permanent Subcommittee on Investigations. *Wall Street and the Financial Crisis: Anatomy of a Financial Collapse* https://www.hsgac.senate.gov//imo/media/doc/Financial_Crisis/FinancialCrisisReport.pdf?attempt=2 (2011).
19. Chon, G., Binham, C. & Noonan, L. Six banks fined $5.6bn over rigging of foreign exchange markets. *Financial Times* (20 May 2015).
20. Capital punishment. *The Economist* (5 July 2014).
21. Wood, R. W. New UBS tax evasion probe, again over Americans. *Forbes* (5 February 2015).
22. Scannell, K., Arnold, M. & Stacey, K. HSBC forex traders charged with criminal fraud. *Financial Times* (20 July 2016).
23. Gray, A., McLannahan, B. & Foley, S. Fake accounts cast cloud over Wells Fargo culture. *Financial Times* (9 September 2016).
24. Samson, A. Wells Fargo hit after auto insurance scandal emerges. *Financial Times* (28 July 2017).
25. Karp, P., Evershed, N. & Knaus, C. A recent history of Australia's banking scandals. *The Guardian* (19 April 2018).
26. *Honesty/Ethics in Professions* http://www.gallup.com/poll/1654/honesty-ethics-professions.aspx (Gallup, accessed 16 June 2016).
27. McCarthy, J. *Americans' confidence in banks still languishing below 30%* http://www.gallup.com/poll/192719/americans-confidence-banks-languishing-below.aspx?g_source=Economy&g_medium=newsfeed&g_campaign=tiles (Gallup, 16 June 2016).
28. Financial Conduct Authority. *Business Plan 2016/17* https://www.fca.org.uk/publication/corporate/business-plan-2016-17.pdf (2016).
29. Australian Securities and Investment Commission. *ASIC's Corporate Plan 2015–2016 to 2018–2019* http://download.asic.gov.au/media/3338908/corporate-plan-2015_published-31-august-2015.pdf (2015).
30. Yong, E. Replication studies: Bad copy. *Nature* **485**, 298–300 (2012).
31. Klein, R. A. et al. Data from investigating variation in replicability: a "many labs" replication project. *J. Open Psychol. Data* **2**, e4 (2014).
32. Klein, R. A. et al. Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci*. **1**, 443–490 (2018).
33. Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
34. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci*. **33**, 61–83 (2010).
35. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Academic, 1969).
36. Gino, F., Ayal, S. & Ariely, D. Self-serving altruism? The lure of unethical actions that benefit others. *J. Econ. Behav. Organ*. **93**, 285–292 (2013).
37. YouGov-Cambridge. *Public Trust in Banking* http://cdn.yougov.com/cumulus_uploads/document/ylf7gpof19/Public_Trust_in_Banking_Final.pdf (2013).
38. Deloitte UK Banking Insight Team. Culture in banking: under the microscope. *The Deloitte Bank Survey 2013* https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-culture-in-banking.pdf (2013).
39. Sieber, J. E., Iannuzzo, R. & Rodriguez, B. Deception methods in psychology: have they changed in 23 years? *Ethics Behav*. **5**, 67–85 (1995).
40. Gerlach, P., Teodorescu, K. & Hertwig, R. The truth about lies: a meta-analysis on dishonest behavior. *Psychol. Bull*. **145**, 1–44 (2019).
41. Abeler, J., Nosenzo, D. & Raymond, D. Preferences for truth-telling. *Econometrica* **87**, 1115–1153 (2019).
42. Rahwan, Z., Hauser, O. P., Kochanowska, E. & Fasolo, B. High stakes: a little more cheating, a lot less charity. *J. Econ. Behav. Organ*. **152**, 276–295 (2018).
43. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav*. **1**, 0021 (2017).
44. Ioannidis, J. P. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
45. Ioannidis, J. P. Why most published research findings are false. *PLoS Med*. **2**, e124 (2005).
46. Lane, D. M. & Dunlap, W. P. Estimating effect size: bias resulting from the significance criterion in editorial decisions. *Br. J. Math. Stat. Psychol*. **31**, 107–112 (1978).
47. Maxwell, S. E., Lau, M. Y. & Howard, G. S. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol*. **70**, 487–498 (2015).
48. Mohan, G. Banking industry culture primes for cheating, study suggests. *Los Angeles Times* (21 November 2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data reporting

It was not possible to control for sample size as this depended on willingness of organizations and staff to participate. We did however, seek out medium-to-large organizations in an effort to get a greater sample size than the original sample, in keeping with best practices for replications. The data were collected using a Qualtrics survey. As such, the randomization process could not be influenced by the researchers and the researchers were blinded to the data collection. We also deleted email addresses provided by participants after the granting of rewards so no (direct) personal identifiers were available, apart from general demographic information.

Experimental design for bankers

Bankers were recruited from two institutions, one in the Asia Pacific ($n = 620$) and one in the Middle East ($n = 148$). The protocol was based on a previously published experimental design[4] and the studies were run in February 2016 and August 2015, respectively.

In summary, each institution invited their staff to participate in an online survey, assuring the confidentiality of their responses. Informed consent was sought from all participants. Once informed consent was granted, participants were randomly allocated to the treatment or control condition. In the control condition, participants were asked questions about their leisure activities before undertaking the coin-flip task. In the treatment condition, participants were asked questions about their profession to activate or 'prime' their professional identity ahead of the task (for example, "Why did you decide to become a bank employee?").

Ahead of undertaking the coin-tossing exercise, participants were informed about the reward mechanism. Specifically, that they could win approximately US$14 for each of the 10 coin tosses, making a maximum potential earning of approximately US$140 in total in the local currency equivalent. Ahead of each coin toss, participants were made aware of the winning outcome. This provided an opportunity for dishonesty.

The experimental design enables comparisons between the reported winning outcomes of the control and treatment groups, and between each of these groups and a binomial distribution ($P = 0.5$). A binomial distribution represents the frequencies of different outcomes that would be expected to emerge for an unbiased or fair coin (that is, 0.5 probability of tossing heads or tails). Although it is impossible to know whether individual participants cheated during the experiment, as they were unobserved, by analysing the aggregated outcomes from both the treatment and control groups, an assessment of dishonesty can be made[49].

Participants were informed that rewards would be calculated by first determining whether their total number of self-reported winning tosses was greater than another randomly drawn participant in the same survey and, if so, the qualifying participant was then entered into a draw in which one in five would win an amount corresponding to the number of winning tosses reported.

The first element introduces a competitive aspect to the reporting of winning coin flips. That is, if a participant expects that their colleagues will over-report the number of winning tosses that they actually experienced, then that participant may be induced to over-report their number of winning tosses to increase the likelihood of receiving a reward. The second element of the reward mechanism—the lottery—was introduced consistent with the original study, to limit the cost of funding the research, and is not expected to affect behaviour[50].

Bankers were informed that any reward would be paid out to them personally in the form of shopping vouchers for popular retail outlets in the respective jurisdictions. Note that in the Asia Pacific sample, an additional 559 participants were able to win a reward for a charity affiliated with the bank, instead of a reward for themselves. The charity was affiliated with the bank, so was expected to be familiar to Asia Pacific bank staff.

Immediately after completing the coin-tossing task, participants were asked for their expectations regarding the number of winning tosses that they expected their colleagues to report on average. This was intended as a direct measure of perceived dishonesty of bankers' peers.

Following the coin-tossing exercise and measure of expectations of peer performance in that exercise, participants were asked to complete a mock investment task. The main purpose of the exercise is to draw attention away from the coin-tossing exercise.

Subsequently, participants were asked to complete a word quiz, in which they had to choose letters to complete words (for example, '_ _ o c k' could be 'clock' or 'stock'). The purpose of this task is to determine whether the priming of professional identity was successful by comparing the number of banking-themed solutions among the control and treatment groups.

Participants then completed questions relating to their various work-related attitudes. The responses were made on a seven-point scale ranging from 'strongly disagree' to 'strongly agree'. The purpose of these questions is to understand participants' relative importance of materialism, competitiveness and self-esteem being determined by others, in general, and with regard to their profession in particular. These assessments are used to explore possible mechanisms for dishonest behaviour.

For the sake of brevity of the survey and given that no significant relationship between risk literacy and frequency of winning tosses was found in the previous study[4], the risk literacy task was removed. Finally, participants were asked a range of questions to obtain personal information (that is, age, gender, education, nationality, experience, income relative to colleagues, type of role and location of role in the bank). For the Asia Pacific study, the question of nationality was excluded as we were advised by the bank that it would be an anomaly for non-nationals to be employed by the bank. We did however add a familiarity check, asking participants: "Are you familiar with research on banking industry culture which uses a survey like the one you just completed?"

Experimental design for non-banking professionals

The protocol for all three non-banking professional studies was based on the experimental design for non-banking professionals from the previous study[4] (see supplementary information section 4.3 of that study[4]). This protocol, in turn, closely follows that used for banking professionals. The key differences broadly pertain to amending references to bankers to encapsulate all types of professionals, and for the Middle Eastern and European samples, amending the manipulation check to refer to regulators of financial services (Supplementary Information 2.2.3). Similar to the original study, we also excluded the risk literacy task.

Asia Pacific. In August 2018, non-banking professionals ($n = 242$) were recruited in the same Asia Pacific jurisdiction as the bankers. The participants were sourced from a professional panel provider. They were screened to ensure that they were currently employed, in either a full-time or part-time capacity, and that they resided in the relevant jurisdiction. The panel provider used a balanced sampling technique to recruit a nationally representative sample with regards to age and gender.

Middle East. Non-banking professionals ($n = 67$) were recruited from a financial services regulator, located in the same country as the Middle Eastern bankers. The survey was conducted in September 2015.

Europe. Non-banking professionals ($n = 205$) were recruited from a financial services regulator, located in Europe. The survey was conducted in January 2016.

Experimental design for experimenter disclosure

This study is motivated by the use of deception[39] in the previous study[4] and our use of incomplete disclosure regarding the stated purpose of the study to participants.

In particular, we wanted to understand what influence variations in experimenter disclosure regarding the purpose of the study had on outcomes from the coin-flipping task. Owing to the largely inaccessible nature of the banking population, we conducted an experiment (n = 925) on Amazon Mechanical Turk (MTurk). We targeted the US general population, as our jurisdictions of interest were understood to have insufficient active participants[51,52] from our previous recruitment experience and running prior surveys.

The London School of Economics Research Ethics Committee gave special dispensation for the use of deception. The survey was conducted on 6 July 2017.

The survey used was modelled on the survey that was used for both bankers and non-bankers as described previously[4]. The manipulation related to the stated experimental purpose. On the landing page of the survey, participants were randomly allocated to transparency (that is, full disclosure); incomplete disclosure; or deception conditions. These conditions corresponded to being told in the introductory statement that it was a study on honesty; norms and attitudes among professionals; or life and satisfaction, respectively. The experimental purpose was repeated on the consent form. All subsequent elements of the survey were the same across the conditions.

After the introductory remarks, participants were requested to provide informed consent. In keeping with the original banking survey[4], they were then asked to complete questions on life satisfaction and leisure activities. The questions on leisure activities formed the control condition in both the banker and non-banker surveys.

Participants were then introduced to the coin-flipping task, in which they could win US 5 cents for each winning outcome, with a maximum of US 50 cents over the 10 rounds. Although this is considerably less than winnings available to bankers and non-bankers, there is growing experimental evidence that the size of the reward has a negligible effect on cheating behaviour, including among the MTurk population from which we sampled[41,42].

In contrast to the original survey[4], there was no uncertainty over whether winnings would relate to either the coin-flipping or mock investment task outcomes, as the mock investment task was excluded for survey brevity. A word puzzle task, work attitudes and demographic questions followed the coin-flipping task. Finally, participants were asked about their experience with coin-flipping tasks and any perceptions of deception in this and other experiments.

Once the survey was closed, all participants were debriefed and for those experiencing deception, an apology was made.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data generated for these studies are available at https://osf.io/56dcp/?view_only=3ef6585039b74bf9aae5deafa0f31e64.

Code availability

The code used for analyses is available at https://osf.io/56dcp/?view_only=3ef6585039b74bf9aae5deafa0f31e64. Please note the code is written in R.

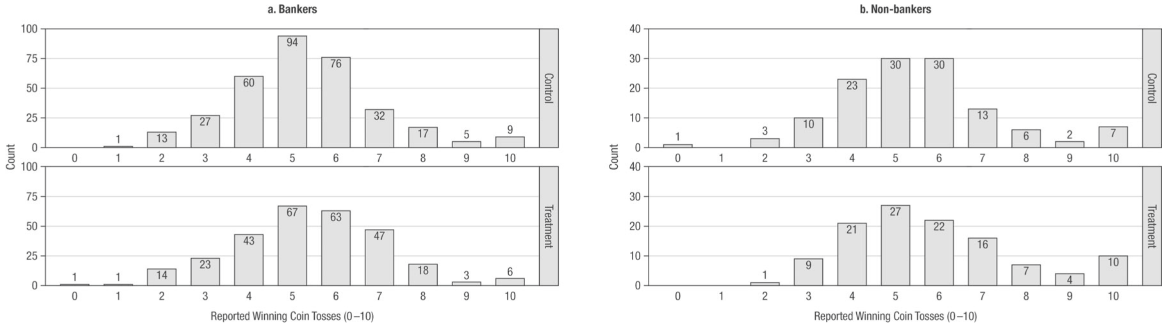
49. Houser, D., Vetter, S. & Winter, J. Fairness and cheating. *Eur. Econ. Rev*. **56**, 1645–1655 (2012).
50. Starmer, C. & Sugden, R. Does the random-lottery incentive system elicit true preferences? An experimental investigation. *Am. Econ. Rev*. **81**, 971–978 (1991).
51. Ross, J., Zaldivar, A., Irani, L. & Tomlinson, B. *Who are the Turkers? Worker demographics in Amazon Mechanical Turk*. Social Code Report 2009-01 (Department of Informatics, University of California, Irvine, 2009).
52. Difallah, D., Filatova, E. & Ipeirotis, P. Demographics and dynamics of Mechanical Turk workers. In *Proc. 11th ACM International Conference on Web Search and Data Mining* 135–143 (ACM, 2018).

Acknowledgements We thank all participating institutions and their staff for joining in this research, E. Kochanowska for her support in conducting additional statistical analyses and N. Obradovich for helpful comments. Our gratitude goes to Google ATAP and their Program Manager G. Pickard for funding this research.

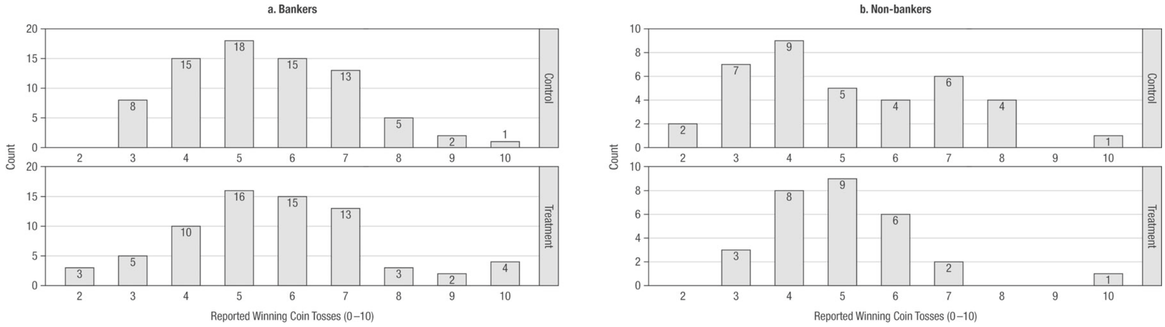
Author contributions Z.R. and B.F. designed the research, Z.R. performed the research and analysed the data and Z.R., E.Y. and B.F. wrote the paper.

Competing interests The authors declare no competing interests.

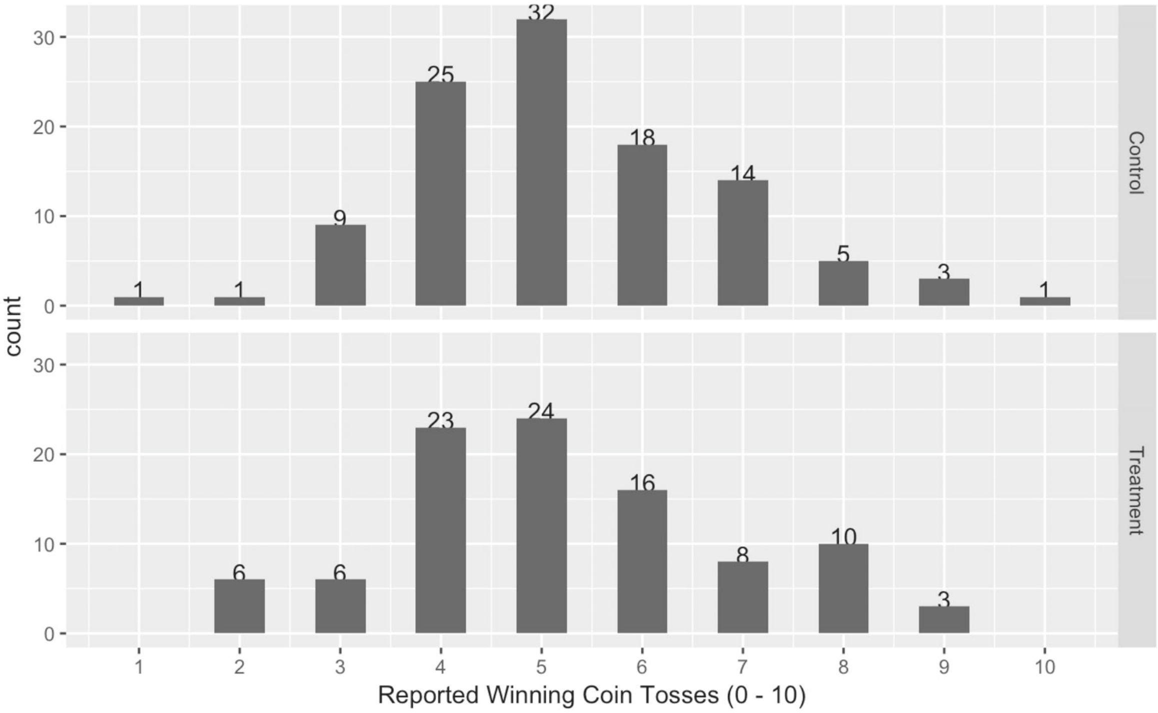
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1741-y.
Correspondence and requests for materials should be addressed to Z.R.
Peer review information *Nature* thanks Bernd Irlenbusch and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



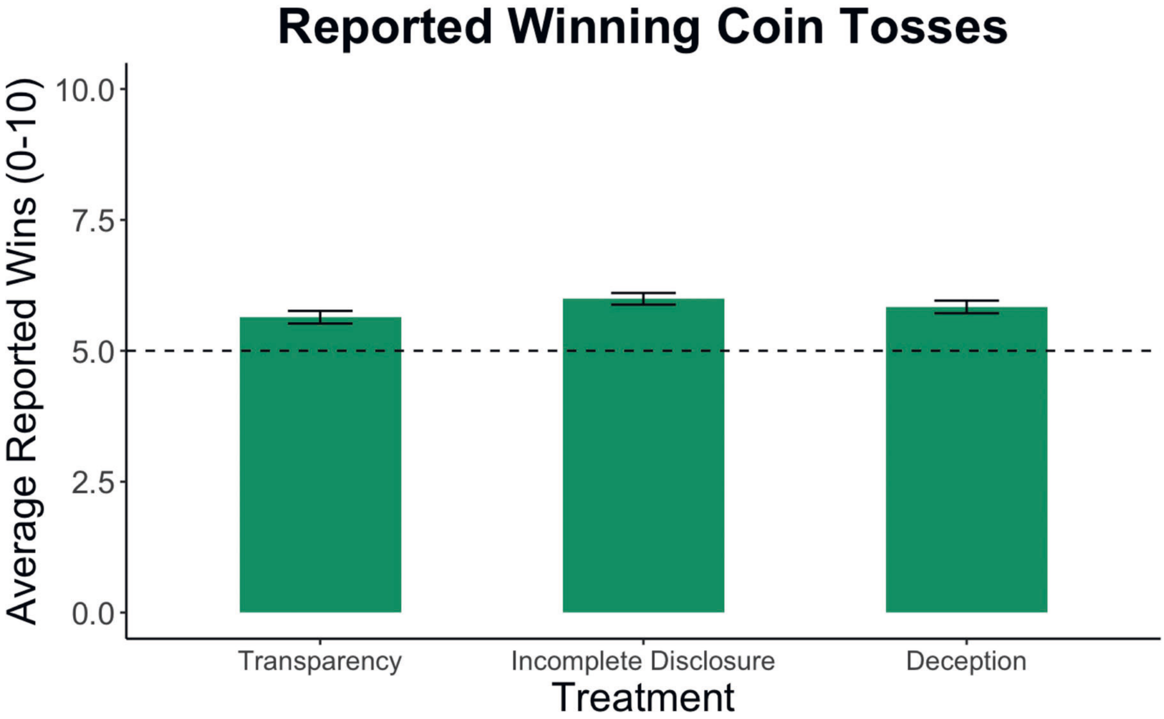
Extended Data Fig. 1 | Distributions of reported winning coin tosses in the Asia Pacific. a, **b**, The frequency of different totals of reported winning coin tosses among Asia Pacific individuals from 10 rounds of the coin-tossing task. **a**, Individuals in the treatment group (*n* = 286) were primed about their professional identity as a banker. Individuals in the control group (*n* = 334) were asked a series of questions about their leisure activities. **b**, Individuals in the treatment group (*n* = 117) were primed about their professional identity. Individuals in the control group (*n* = 125) were asked the same series of questions on leisure activities as the bankers. Individuals reporting to be currently in banking roles were excluded.



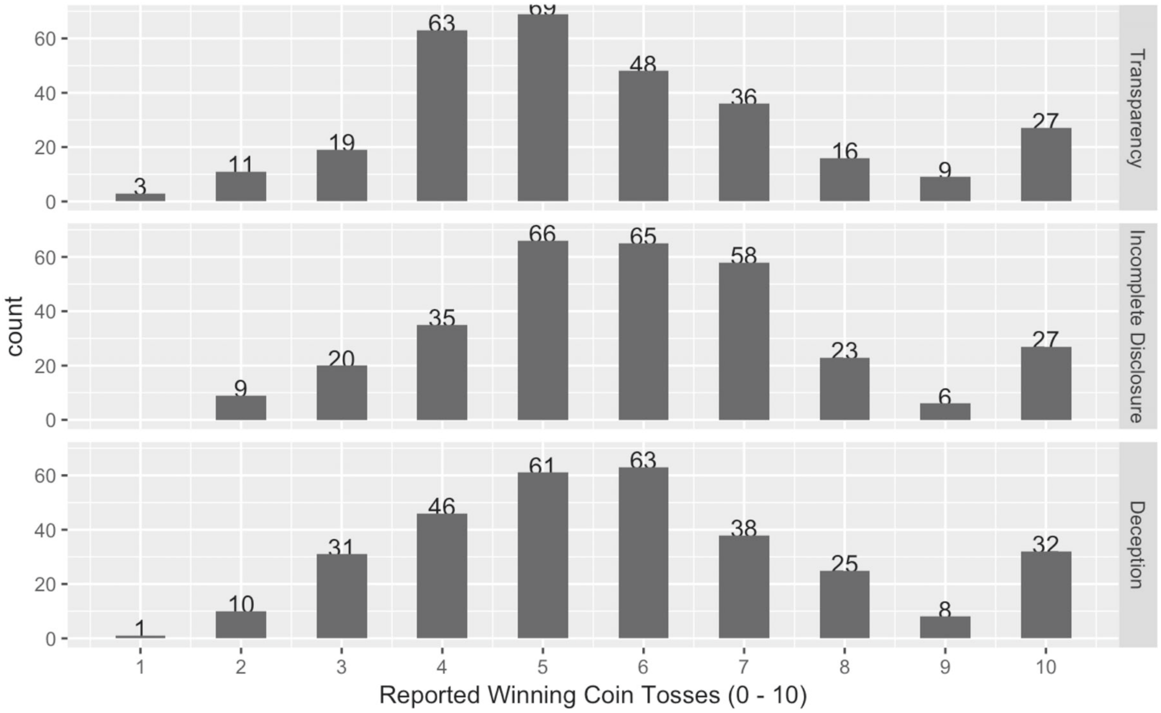
Extended Data Fig. 2 | Distributions of reported winning coin tosses in the Middle East. a, **b**, The frequency of different totals of reported winning coin tosses among Middle Eastern individuals from 10 rounds of the coin-tossing task. **a**, Individuals in the treatment group (*n* = 71) were primed about their professional identity as a banker. Individuals in the control group (*n* = 77) were asked a series of questions about their leisure activities. **b**, Individuals in the treatment group (*n* = 29) were primed with their regulatory identity in financial services. Individuals in the control group (*n* = 38) were asked the same series of questions on leisure activities as the bankers.



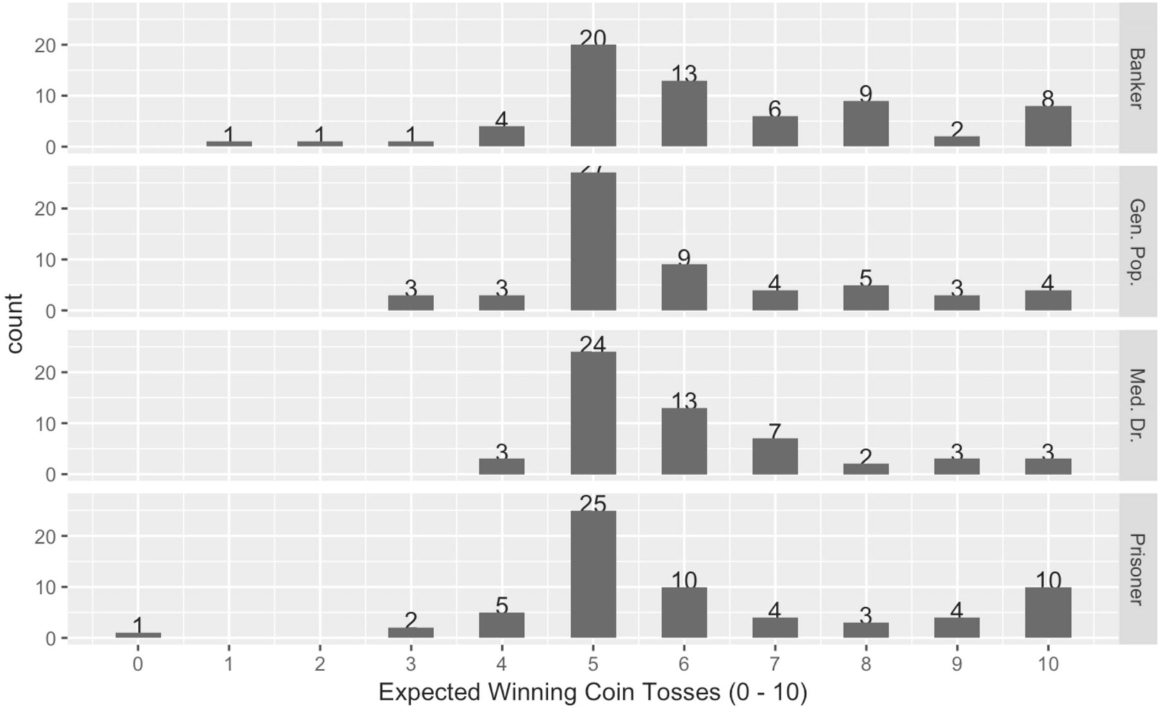
Extended Data Fig. 3 | Distributions of reported winning coin tosses of regulators of financial services (non-bankers) in Europe. The frequency of different totals of reported winning coin tosses among European regulators from 10 rounds of the coin-tossing task. Individuals in the treatment group ($n = 96$) were primed with their regulatory identity in the financial services industry. Individuals in the control group ($n = 109$) were asked a series of questions about their leisure activities.



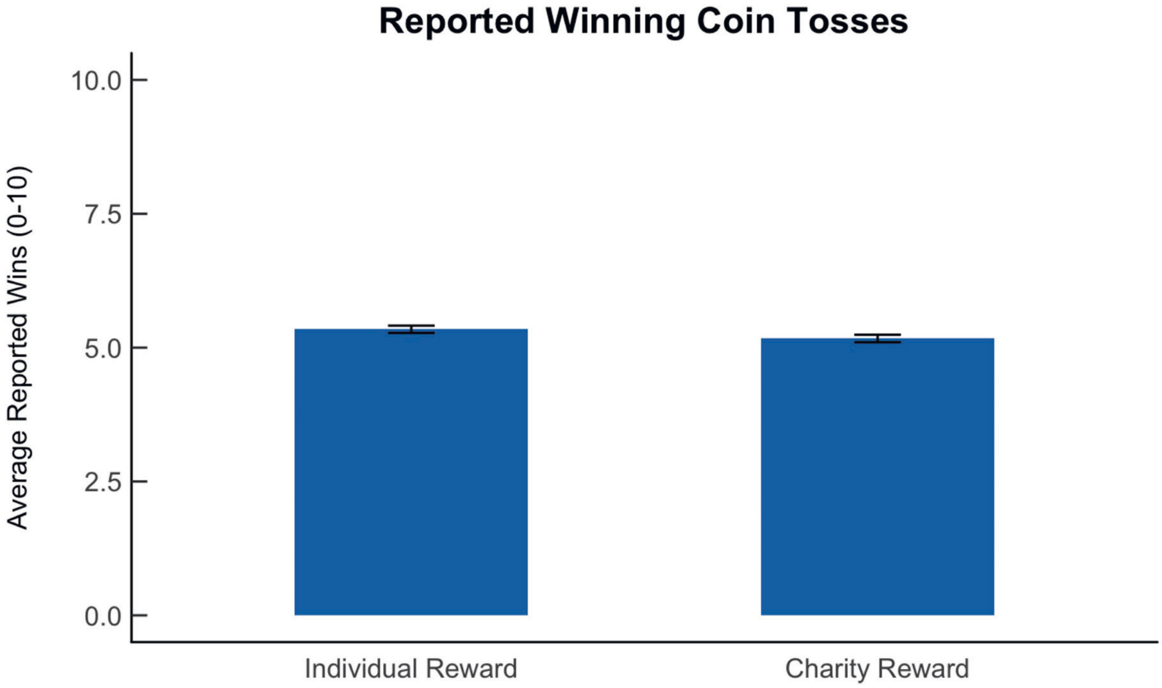
Extended Data Fig. 4 | Effect of experimenter disclosure on honesty. The average number of winning coin tosses out of 10 reported by participants in each of the three conditions (n = 925) that varied the disclosed purpose of the experiment: deception (life and satisfaction), incomplete disclosure (norms and attitudes among professionals) and transparency (honesty). Data are mean ± s.e.m. No differences were found between deception (M = 5.84)—as used previously[4]—and incomplete disclosure (M = 5.99), which was used here (P = 0.188, one-tailed Wilcoxon rank-sum test, α-adjusted for family-wise errors: 0.05/3 = 0.017). The only statistical difference found between the conditions was that those in the incomplete disclosure condition reported a higher number of winning coin tosses than those in the transparency condition (M = 5.64, P = 0.006, one-tailed Wilcoxon rank-sum test). The difference between the average outcomes of these two conditions was negligible and not sufficient to change actual pay-offs for participants.



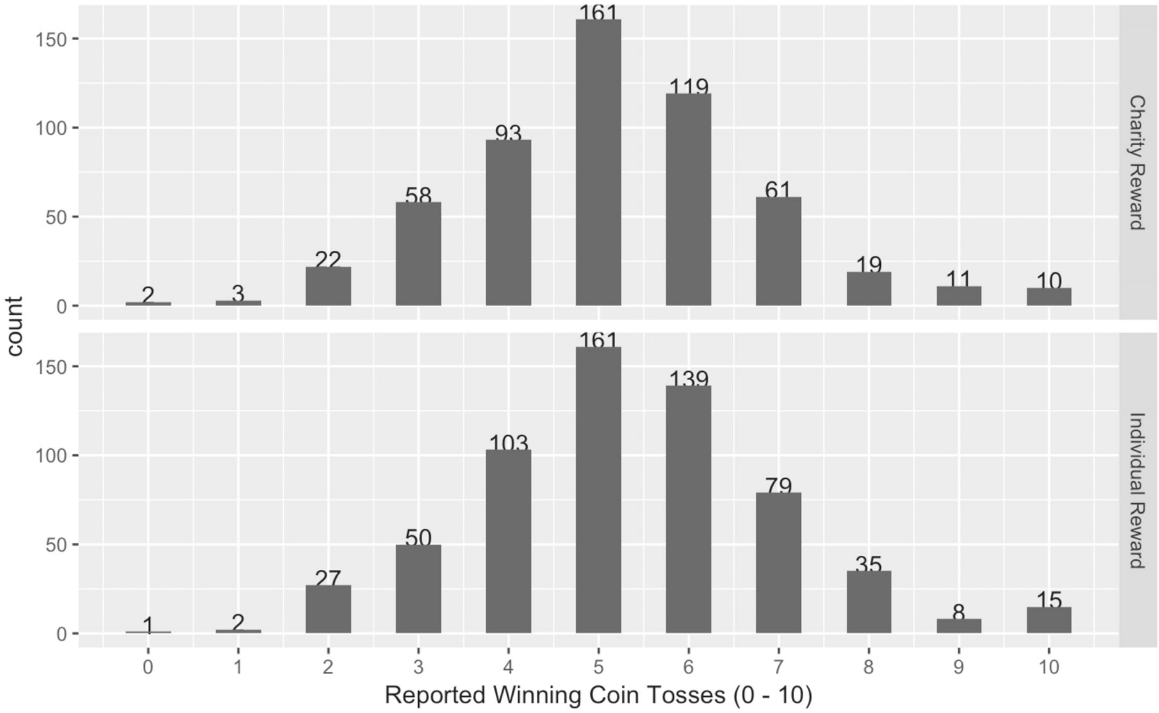
Extended Data Fig. 5 | Underlying distributions of the effect of experimenter disclosure on honesty. The frequency of different totals of reported winning coin tosses from 10 rounds of a coin-tossing task among MTurk participants. In this experiment, the disclosed purpose of the experiment was randomly assigned to be in one of three conditions; deception, in which participants were informed that they were in a study regarding life and satisfaction ($n = 315$), incomplete disclosure, in which participants were informed that they were in a study regarding the norms and attitudes among professionals ($n = 309$) and transparency, in which participants were informed that we were studying honesty ($n = 301$).



Extended Data Fig. 6 | Underlying distributions of expectations of the honesty of others. The expected frequency of different totals of reported winning coin tosses from 10 rounds of a coin-tossing task among a sample of Asia Pacific non-banking professionals, sourced from a panel. In this experiment, the participants themselves had experience of the coin-tossing task ahead of being questioned on their expectations of reported winning outcome from one of four different populations. As such, participants had the opportunity to learn that one could be dishonest in the task. Participants were randomly assigned to be asked expectations of reported winning outcomes for the following populations: bankers ($n = 65$), general population ($n = 58$), medical doctors ($n = 55$) and prison inmates ($n = 64$).

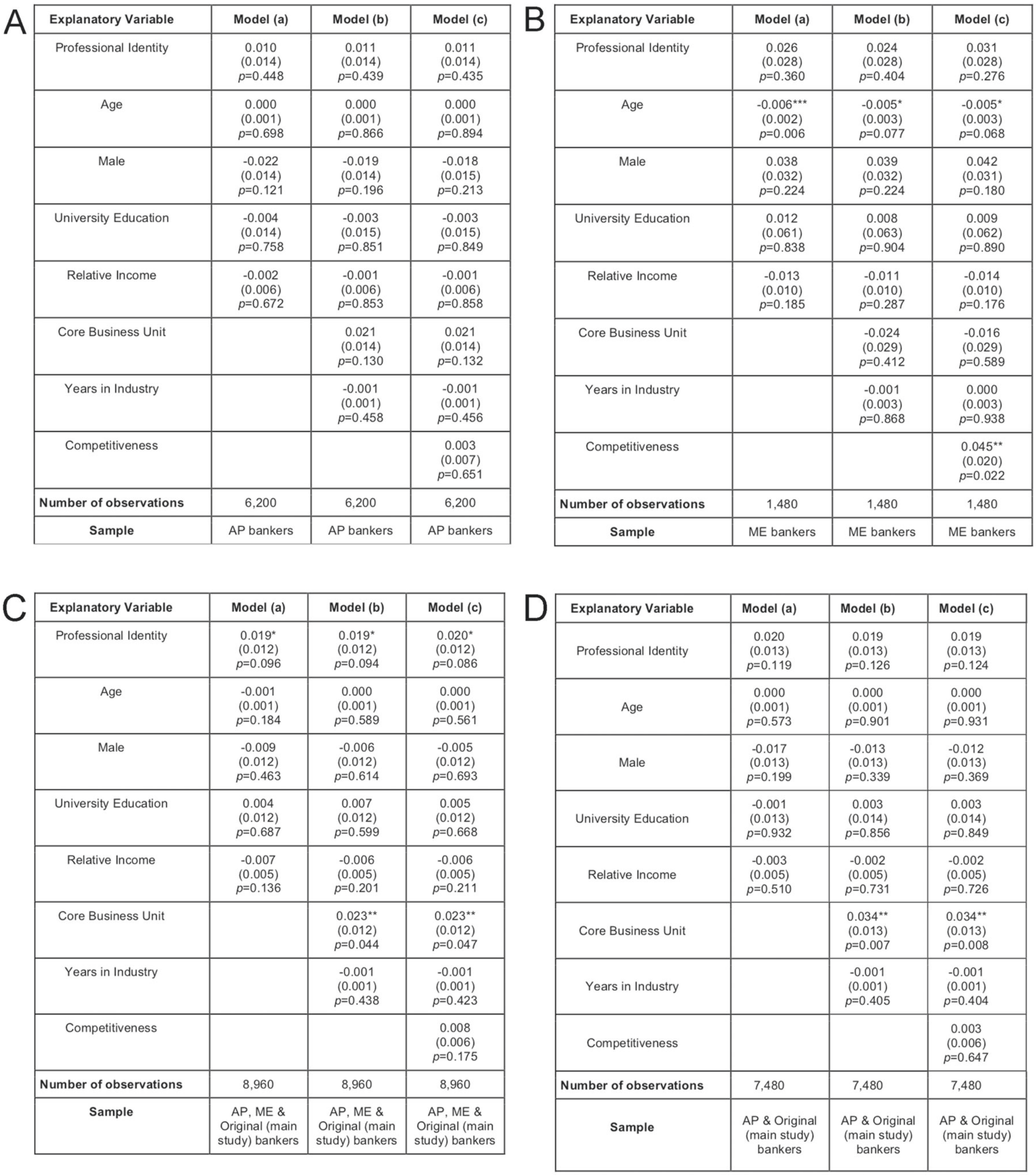


Extended Data Fig. 7 | Effect of the nature of the reward on honesty. The average number of winning coin tosses out of 10 reported by Asia Pacific bankers in conditions in which they can either win money for themselves ($n = 620$) or charity ($n = 559$). Data are mean ± s.e.m. On average, those winning money for themselves and for charity reported 5.34 and 5.17 winning tosses, respectively. No difference was found between those able to win money—up to around US$140—for themselves or charity ($P = 0.073$, two-tailed Wilcoxon rank-sum test).



Extended Data Fig. 8 | Underlying distributions of the effect of the nature of the reward on honesty. The frequency of different totals of reported winning coin tosses from 10 rounds of a coin-tossing task among Asia Pacific bankers. Participants were able to either win a reward for a charity ($n = 559$) or for themselves ($n = 620$).

Extended Data Table 1 | Effect of professional identity on honesty among bank employees



A

Explanatory Variable	Model (a)	Model (b)	Model (c)
Professional Identity	0.010 (0.014) *p*=0.448	0.011 (0.014) *p*=0.439	0.011 (0.014) *p*=0.435
Age	0.000 (0.001) *p*=0.698	0.000 (0.001) *p*=0.866	0.000 (0.001) *p*=0.894
Male	-0.022 (0.014) *p*=0.121	-0.019 (0.014) *p*=0.196	-0.018 (0.015) *p*=0.213
University Education	-0.004 (0.014) *p*=0.758	-0.003 (0.015) *p*=0.851	-0.003 (0.015) *p*=0.849
Relative Income	-0.002 (0.006) *p*=0.672	-0.001 (0.006) *p*=0.853	-0.001 (0.006) *p*=0.858
Core Business Unit		0.021 (0.014) *p*=0.130	0.021 (0.014) *p*=0.132
Years in Industry		-0.001 (0.001) *p*=0.458	-0.001 (0.001) *p*=0.456
Competitiveness			0.003 (0.007) *p*=0.651
Number of observations	6,200	6,200	6,200
Sample	AP bankers	AP bankers	AP bankers

B

Explanatory Variable	Model (a)	Model (b)	Model (c)
Professional Identity	0.026 (0.028) *p*=0.360	0.024 (0.028) *p*=0.404	0.031 (0.028) *p*=0.276
Age	-0.006*** (0.002) *p*=0.006	-0.005* (0.003) *p*=0.077	-0.005* (0.003) *p*=0.068
Male	0.038 (0.032) *p*=0.224	0.039 (0.032) *p*=0.224	0.042 (0.031) *p*=0.180
University Education	0.012 (0.061) *p*=0.838	0.008 (0.063) *p*=0.904	0.009 (0.062) *p*=0.890
Relative Income	-0.013 (0.010) *p*=0.185	-0.011 (0.010) *p*=0.287	-0.014 (0.010) *p*=0.176
Core Business Unit		-0.024 (0.029) *p*=0.412	-0.016 (0.029) *p*=0.589
Years in Industry		-0.001 (0.003) *p*=0.868	0.000 (0.003) *p*=0.938
Competitiveness			0.045** (0.020) *p*=0.022
Number of observations	1,480	1,480	1,480
Sample	ME bankers	ME bankers	ME bankers

C

Explanatory Variable	Model (a)	Model (b)	Model (c)
Professional Identity	0.019* (0.012) *p*=0.096	0.019* (0.012) *p*=0.094	0.020* (0.012) *p*=0.086
Age	-0.001 (0.001) *p*=0.184	0.000 (0.001) *p*=0.589	0.000 (0.001) *p*=0.561
Male	-0.009 (0.012) *p*=0.463	-0.006 (0.012) *p*=0.614	-0.005 (0.012) *p*=0.693
University Education	0.004 (0.012) *p*=0.687	0.007 (0.012) *p*=0.599	0.005 (0.012) *p*=0.668
Relative Income	-0.007 (0.005) *p*=0.136	-0.006 (0.005) *p*=0.201	-0.006 (0.005) *p*=0.211
Core Business Unit		0.023** (0.012) *p*=0.044	0.023** (0.012) *p*=0.047
Years in Industry		-0.001 (0.001) *p*=0.438	-0.001 (0.001) *p*=0.423
Competitiveness			0.008 (0.006) *p*=0.175
Number of observations	8,960	8,960	8,960
Sample	AP, ME & Original (main study) bankers	AP, ME & Original (main study) bankers	AP, ME & Original (main study) bankers

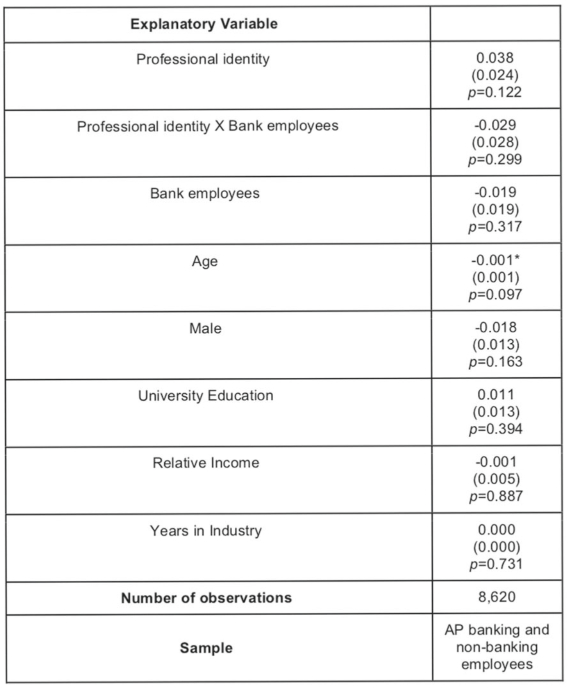
D

Explanatory Variable	Model (a)	Model (b)	Model (c)
Professional Identity	0.020 (0.013) *p*=0.119	0.019 (0.013) *p*=0.126	0.019 (0.013) *p*=0.124
Age	0.000 (0.001) *p*=0.573	0.000 (0.001) *p*=0.901	0.000 (0.001) *p*=0.931
Male	-0.017 (0.013) *p*=0.199	-0.013 (0.013) *p*=0.339	-0.012 (0.013) *p*=0.369
University Education	-0.001 (0.013) *p*=0.932	0.003 (0.014) *p*=0.856	0.003 (0.014) *p*=0.849
Relative Income	-0.003 (0.005) *p*=0.510	-0.002 (0.005) *p*=0.731	-0.002 (0.005) *p*=0.726
Core Business Unit		0.034** (0.013) *p*=0.007	0.034** (0.013) *p*=0.008
Years in Industry		-0.001 (0.001) *p*=0.405	-0.001 (0.001) *p*=0.404
Competitiveness			0.003 (0.006) *p*=0.647
Number of observations	7,480	7,480	7,480
Sample	AP & Original (main study) bankers	AP & Original (main study) bankers	AP & Original (main study) bankers

The dependent variable is a reported winning toss. The reported results are marginal effects calculated at the median levels of the covariates, and the standard errors (in parentheses) have been corrected for clustering at the individual level. The median covariates are a measure of the change in probability of reporting a winning outcome. For each of the following samples and pools of samples, model a shows reported winning tosses that are regressed on a dummy variable for the professional identity condition and individual characteristics. Model b extends model a to include work-related variables. Model c extends model b to include an additional control of self-reported materialism. These models are drawn from the original study[4]. **a**, Probit estimates for the Asia Pacific (*n* = 620). **b**, Probit estimates for the Middle East (*n* = 148). **c**, Probit estimates for pooled data from the Asia Pacific (*n* = 620), Middle East (*n* = 148) and original main study[4] (*n* = 128), resulting in *n* = 896 participants. **d**, Probit estimates for pooled data from the Asia Pacific (*n* = 620) and original main study[4] (*n* = 128)—the samples for which the manipulation check worked—resulting in a combined *n* = 748 participants. **P* < 0.10, ***P* < 0.05, ****P* < 0.01 (two-tailed Wald tests).

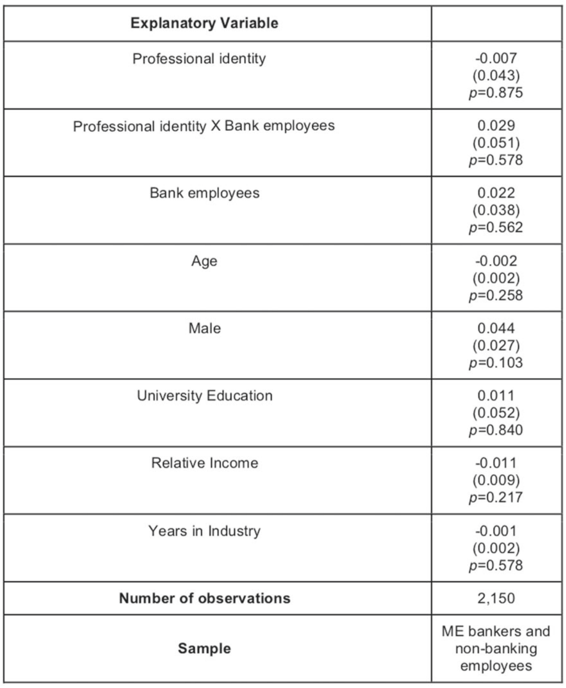
Extended Data Table 2 | Effect of professional identity on honesty in bank and non-banking employees

A



Explanatory Variable	
Professional identity	0.038 (0.024) *p*=0.122
Professional identity X Bank employees	-0.029 (0.028) *p*=0.299
Bank employees	-0.019 (0.019) *p*=0.317
Age	-0.001* (0.001) *p*=0.097
Male	-0.018 (0.013) *p*=0.163
University Education	0.011 (0.013) *p*=0.394
Relative Income	-0.001 (0.005) *p*=0.887
Years in Industry	0.000 (0.000) *p*=0.731
Number of observations	8,620
Sample	AP banking and non-banking employees

B



Explanatory Variable	
Professional identity	-0.007 (0.043) *p*=0.875
Professional identity X Bank employees	0.029 (0.051) *p*=0.578
Bank employees	0.022 (0.038) *p*=0.562
Age	-0.002 (0.002) *p*=0.258
Male	0.044 (0.027) *p*=0.103
University Education	0.011 (0.052) *p*=0.840
Relative Income	-0.011 (0.009) *p*=0.217
Years in Industry	-0.001 (0.002) *p*=0.578
Number of observations	2,150
Sample	ME bankers and non-banking employees

The dependent variable is a reported winning toss. The reported results are marginal effects calculated at the median levels of the covariates, and the standard errors (in parentheses) have been corrected for clustering at the individual level. The median covariates are a measure of the change in probability of reporting a winning outcome. **a**, Probit estimates for Asia Pacific. The model is as described previously, run on participants of the original main study[4] (*n* = 128) and the non-banker population (*n* = 133) to demonstrate that priming professional identity led to higher levels of reported winning tosses relative to non-bankers. Reported winning tosses are regressed on a dummy for the professional identity condition and individual characteristics, and an interaction term for professional identity and bank employees (*n* = 620 (bankers) + 242 (non-bankers) = 862). **b**, Probit estimates for the Middle East. The model is as described previously, run on participants of the original main study[4] (*n* = 128) and the non-banker population (*n* = 133) to demonstrate that priming professional identity led to higher levels of reported winning tosses relative to non-bankers. Reported winning tosses are regressed on a dummy for the professional identity condition and individual characteristics, and an interaction term for professional identity and bank employees (*n* = 148 (bankers) + 67 (non-bankers) = 215). **P* < 0.10, ***P* < 0.05, ****P* < 0.01 (two-tailed Wald tests).

nature research

Corresponding author(s): Zoe Rahwan

Last updated by author(s): Jan 31, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Qualtrics was used to execute the all surveys (i.e. for bankers, non-bankers and MTurkers).
Data analysis	Data analysis was conducted using R (Version 1.0.153)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data and code for all statistical tests and analyses will be publicly available either through OSF and/or the Nature website. A separate file has been generated for the raw data presented in the figures in the main article, to be made available on the Nature website.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Data are quantitative, gathered in 5 field studies and one on-line (MTurk) study.
Research sample	1) Bankers in the Asia Pacific, recruited from the one institution 2) Bankers in the Middle East, recruited from the one institution, 3) Non-bankers in the Asia Pacific were sought via a panel provider and to be representative on age and gender, 4) Non-bankers in the Middle East, recruited from the one institution, 5) Non-bankers in Europe, recruitment from the one institution, and 6) MTurkers, recruited in the US.
Sampling strategy	Being field studies requiring organizational permission, strictly speaking studies 1,2,4,5 are convenience samples, and as such the sample size could not be pre-determined. Our power analysis calculations however provided an estimate of the needed sample size to achieve the effect of the original study we were attempting to replicate. In study 3, we strove to recruit a nationally representative sample on gender and age attributes, around double the size of the original sample and well above the size needed to identify the treatment effect found in the original banker study. In Study 6, based on previous studies, we sought a sample size to detect a difference of an average 0.5 winning coin tosses out of 10 rounds between conditions, with 90% power and a significance level of 5%.
Data collection	The data was collected via the Qualtrics survey. As such, the randomization process could not be influenced by the researchers and the researchers were blind to the data collection.
Timing	Study 1: February 2016, Study 2: August 2015, Study 3: August 2018, Study 4: September 2015, Study 5: January 2016, Study 6: July 2017
Data exclusions	For Study 3 - a non-banker study - participants who reported being bankers were excluded from the final data set. Robustness checks were conducted for those who reported previously having worked in the banking sector.
Non-participation	Incomplete surveys as a % of all surveys: 1) 42, 2) 31, 3) 70, 4), 26, 5) 56, 6) 0. Approximated participation rates (completed surveys as a % of total staff): 1) 16, 2) 6, 3) n/a, 4) 50, 5) 6, 6) n/a (Mturk)
Randomization	Randomization was undertaken in all studies by Qualtrics.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above and SI for descriptive statistics for Studies 1, 2, 3, 4 & 5.
Recruitment	Participants for studies #1,2,4,5 were recruited through contacts in the financial services industry around the world. Our experience informs us that institutional selection bias is a nearly insurmountable challenge in field replications of high profile work, as discussed in the manuscript. We don't identify issues with (self-reported) familiarity with the previous study and survey technique. As for Study #3, we recruited panel participants through a professional panel services provider, Qualtrics. For Study #6, we recruited participants through Amazon Mechanical Turk.
Ethics oversight	LSE Research Ethics Committee. Reference Number 000582 (now fully referenced in the SI)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

nature

https://doi.org/10.1038/s41586-019-1741-y

Supplementary information

Heterogeneity in banker culture and its influence on dishonesty

In the format provided by the authors and unedited

Zoe Rahwan, Erez Yoeli & Barbara Fasolo

Supplementary Information

Contents

1.1. Differences with the original banker survey

Due to practical constraints, our studies differed slightly from Cohn et al's. We now describe these differences and postulate their impact.

In introducing the survey, we used incomplete disclosure rather than deception, regarding the purpose of the survey. We draw on the definition of deception that participants "may be given, or be caused to hold, false information about the main purpose of the experiment."[39] Participants in our studies were informed that 'norms and attitudes among professionals' were being researched, but not that of honesty explicitly. By contrast, the original study purported to be about 'Life and Satisfaction among Employees.' Such deception was not deemed necessary for the purposes of the research and had compromised possible institutional participation and ethics approval by the Research Ethics Committee of the London School of Economics and Political Sciences. This approach may have made the participants more self-aware and thus less likely to engage in dishonest behaviour. A recent meta-analysis of honesty studies found that the use of deception by experimenters in disclosing the purpose of a study was actually associated with less rather than more dishonesty.[44] Further, our own experiments reveal no significant differences in dishonesty between participants in incomplete disclosure and deception conditions (SI 1.5).

The potential total value of the reward in both studies was USD 140 - below the USD 200 equivalent used in the original study. The decision to use a lesser amount was guided by consideration of local currency denominations in granting rewards. Despite being less than the original reward, it still provides a substantial financial incentive to over-report the number of winning tosses, especially given that the survey was communicated to take approximately ten minutes. Nevertheless, it is unclear the impact of the lesser financial reward. Traditional economic theory[53] would predict lesser cheating. However there is growing evidence that individuals are largely insensitive to rewards.[40,41] Moreover, some researchers proposing a psychologically-based theory of dishonesty find modest evidence that higher stakes can curtail dishonesty.[54]

The competitive mechanism used to distribute rewards also differed. Cohn et al[4] made a comparison with a participant drawn from a pilot study. Given the limited time and opportunity to run experiments, the protocol for this research was amended to undertake the comparison with a randomly drawn participant from the same study. Given the reduced psychological distance to the benchmark, this would be expected to exacerbate any effect from the expectations of peer behaviour so it is not problematic. Further, in an unintended error, bankers were only entered into the draw to win their earnings if the total number of winning tosses was greater than a colleagues' rather than greater than or equal to as in the original study. This difference would be expected to result in a greater treatment effect given the competitive nature of banking so again we do not believe this would account for smaller effect sizes.

Unlike the original study, the order of the coin-flipping and investment tasks were not randomised because Cohn et al found no effect from task order in their study. We placed the coin-flipping task ahead of the risk-taking task.

1.2. Differences with the Original Non-banker Survey

Similar to the AP banker sample, for all three non-banker studies we used incomplete disclosure in describing the purpose of the study. Again, participants were informed that ‘norms and attitudes among professionals’ were being researched. As discussed in SI 1.1 and in presenting our experimental results in SI 1.5, we have diminished concerns that this affected the outcome of our results relative to Cohn et al’s studies, all of which deployed deception.

The level of rewards were also the same as the bankers in the Asia Pacific and Middle Eastern jurisdictions: ~USD 14 per coin toss, marking a maximum reward of ~USD 140. This is less than the ~USD 20 per coin toss rewards used in the original study. This is unlikely to drive any difference in reported winning outcomes in light of previous meta-analytical finds of honesty being largely insensitive to rewards,[40] even at elevated levels for some populations.[41] In terms of field evidence, we find in the AP banker survey that, when charitable rewards were made available - i.e. winnings of up to ~USD 140 would be donated to a charity affiliated with the bank - the level of reported winning outcomes was indistinguishable from when bankers could win rewards for themselves (p=0.073, two-sided rank-sum test, Extended Data Figure 7).

In the soft launch of the AP non-banker survey, we identified that at least one banker was in the sample based on their responses to the professional identity priming questions. To ensure the absence of any bankers participating in the survey, for the full launch of the survey we asked in the demographics section of the survey whether participants had any experience in the banking sector, to identify any bankers in the control group. If they answered ‘yes’, we then asked how many years of experience they had in banking. The full sample for this population (n=242) excluded all respondents in both control and treatment groups who identified banking as their current profession (n=6) and those in the control group from the soft launch for whom their profession was unknown (n=11). These amendments to the survey and subsequent exclusions increased the confidence with which we could describe the sample as representing non-bankers.

In the AP non-banker study, we also sought expectations from the non-bankers regarding the behaviour of other groups. Again, we drew on the questions posed from Cohn et al[4], asking about expectations for reported winning outcomes in the coin-flipping task for (i) bankers, (ii) prisoners, (iii) doctors and, (iv) the general population and used a between-subject design. In contrast to Cohn et al, we asked these questions to the balanced sample of non-banking professionals, rather than a convenience sample of the general population (n=183, males only). We believe this improves representativeness of the findings. These questions were placed after the coin-flipping task, so the participants had the opportunity to learn of the possibility to cheat, and just ahead of the demographic questions, so as to not influence the outcomes of the mock investment task.

2. Supplementary Data

2.1. Banker Data

2.1.1. Descriptive statistics

	Cohn et al		Asia Pacific Bankers		Middle East Bankers	
	Control	Treatment	Control	Treatment	Control	Treatment
Participants (number)	67	61	334	286	77	71
Age (average years)	39.2	38.5	43.6	42.2	35.4	36.5
Males (proportion - %)	59	63	39	41	46	51
Professional Experience (average years)	12.0	10.9	15.3	15.5	10.5	10.3
University Education (proportion - %)	66	57	37	41	90	96
Core Business Unit (proportion - %)	45	52	51	53	55	46

2.1.2. Randomization checks

Bankers were randomly and evenly allocated between treatment and control conditions by the survey platform, Qualtrics. Non-parametric tests were used to assess whether proper randomization had been achieved. Wilcoxon rank sum tests (two-sided) were used for interval variables, while Chi-squared tests (one-sided) were used for binomial variables. The estimates are broadly consistent with randomization being achieved. The only marginal difference was found with regard to age in the treatment and control groups among the Asia Pacific bankers.

a. Asia Pacific Bankers

	Total Sample (n=620 participants)		Control (n=334 participants)		Treatment (n=286 participants)		
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	p-value
Age (years)^	42.9	0.619	43.6	0.663	42.2	0.453	0.094*
Male	0.40	0.020	0.386	0.027	0.413	0.029	0.504
Professional Experience (years)	15.4	0.457	15.3	0.603	15.5	0.700	0.882

University Education	0.390	0.020	0.374	0.027	0.409	0.029	0.376
Core Business Unit	0.521	0.020	0.515	0.027	0.528	0.030	.747

b. Middle Eastern Bankers

	Total Sample (n=148 participants)		Control (77 participants)		Treatment (71 participants)		
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	p-value
Age (years)^	36.0	7.13	35.4	6.97	36.5	7.30	0.213
Male	0.71	0.46	0.70	0.46	0.72	0.45	0.821
Professional Experience (years)	10.4	6.31	10.4	6.45	10.3	6.19	0.785
University Education	0.93	0.26	0.90	0.31	0.96	0.20	0.155
Core Business Unit	0.51	0.50	0.55	0.50	0.47	0.50	0.328

* p-value <0.10 ^ For each of the 10-year age brackets, the mid-point was used as a proxy for the age (e.g. if a participant was in the 20-29 age bracket, 25 was used as a proxy).

2.1.3. Priming checks

Bankers were asked to solve six word puzzles, four of which had banking-themed solutions. Comparisons of a number of banking-themed solutions between the control and treatment group is used to assess the effectiveness of priming professional identity.

a. Asia Pacific Bankers

The word puzzles presented to the Asia Pacific bankers were amended from Cohn et al given indications from the Middle Eastern bankers of effects from being a commercial rather than investment bank, and cross-cultural differences. The following banking-themed word puzzles were used:
_ _ v i n g -> saving
c r _ _ _ _ -> credit
_ o n e y -> money
_ _ _ n c h -> branch

The number of puzzles solved with banking words was significantly different between the treatment and control group (mean$_{\text{treatment}}$ = 1.7, mean$_{\text{control}}$ = 1.4, rank-sum test (one-sided), Z=-3.021, p-value=0.001), indicating that the priming of professional identity "worked" and assisted in those treated solving more puzzles with banking-themed solutions.

b. Middle Eastern Bankers

The puzzles solved by the Middle Eastern bankers was drawn directly from Cohn et al. They included the following banking-themed puzzles.

_ _ o c k -> stock
_ _ o k e r -> broker
_ o n e y -> money
B _ n d -> bond

The number of puzzles solved with banking-themed words was not significantly different between the treatment and control group (mean$_{treatment}$ = 1.8, mean$_{control}$ = 1.7, rank-sum test (one-sided), Z=-0.706, p-value=0.240), indicating that the priming of professional identity was not effective, unlike in the original study. However, given the similarity in experimental protocol, there is no obvious reason why the priming itself would be ineffective. Rather, it may be that the mechanism used to assess the manipulation was flawed due to unanticipated differences in local vernacular. For example the ‘broker’ puzzle was commonly solved with ‘cooker’ - a colloquial word for ‘stove’ or ‘oven’ in the region. Further, Cohn et al’s study[4] was conducted at an investment bank, whereas the Middle Eastern institution is a commercial bank. This may explain that only two of 71 subjects in the treatment group solved the ‘_ _ o c k’ word puzzle with ‘stock’ while the bulk (47) chose ‘clock.’ Given these considerations, it may be the case that the priming was effective though the measure to test it was flawed.

2.1.4. Fair coin comparisons

As the probability of tossing a winning outcome of a fair coin is known - 0.5 on average - a distribution of winning outcomes from a fair coin over 10 rounds can be calculated. We compare the reported outcomes from our experimental samples with the theoretical distribution of a fair coin to determine if there is any apparent dishonesty. All test results reported below pertain to one-sided rank-sum tests, based on the hypotheses that bankers will cheat more than a fair coin would predict.

a. Asia Pacific Bankers

The treatment group and the control group both report higher winning outcomes than are predicted by a fair coin (p =0.002 and p =0.029, respectively).

b. Middle Eastern Bankers

The treatment group reports higher winning outcomes than are predicted by a fair coin (p =0.017), while the amount of cheating only tended to be significant for the control group (p =0.051).

2.1.5. Familiarity checks

In the AP banker study, we attempted to measure familiarity with the survey tool used. At the end of the survey, we asked the question: Are you familiar with research on banking industry culture which uses a survey like the one you just completed?

Approximately 30% of the respondents answered 'yes.' We found that reporting familiarity with the study could not account for the smaller effect found from priming. Among those that reported being familiar with the study, there is no statistical difference between the winning tosses they reported, whether or not their professional identity was made salient ($\text{mean}_{\text{treated \& familiar}}$ = 5.6, $\text{mean}_{\text{control \& familiar}}$ = 5.5, p=0.160, rank-sum (one-sided)). Similarly, among the AP bankers reporting being unfamiliar with the study, there is no difference in the reported winning outcomes among the primed bankers and those in the control group ($\text{mean}_{\text{treated \& unfamiliar}}$ = 5.3, $\text{mean}_{\text{control \& unfamiliar}}$ = 5.2, p=0.159, rank-sum test (one-sided)).

2.1.6. Channels for Dishonesty

Cohn et al[4] explore various channels through which the salience of professional identity may have increased dishonest behaviour, including materialism, norm obedience and competitiveness (as measured by both self-reports and that associated with varying compensation structures). Given our smaller effect sizes to the extent they become statistically indistinguishable from zero, one would not expect to find evidence for those channels in our studies, and, indeed, we generally did not. We expanded the original study to better understand norm obedience and found some, though not robust, evidence that priming banker identity did change reported expectations of peer behaviour (SI 2.1.6b*)*.

a. Materialism

Materialism, as self-reported on a 1-7 scale, is the only channel for which the authors of the original study find direct empirical support of, noting greater materialism (p=0.017, rank-sum test (one-sided), n=128) among primed bankers and that it is correlated with greater dishonesty (spearman's $\text{rho}_{\text{two-tailed}}$=.237, p=0.007). We do not find support for this among our samples, with materialism not increasing significantly with professional identity priming, either among individual samples (AP bankers: p=0.575, ME bankers: p=0.118, rank-sum test (one-sided), n=620 and n=148, respectively) or among AP and ME bankers combined (p=0.335). However, if we extend the sample to include Cohn et al's bankers from the main study, materialism among primed bankers approaches marginal significance ($\text{mean}_{\text{treatment}}$ = 4.44, $\text{mean}_{\text{control}}$ =4.29, p=0.111) and the correlation with reporting winning coin tosses is weak (spearman's $\text{rho}_{\text{two-tailed}}$ =.077, p=0.021).

b. Norm obedience

Cohn et al[4] also explore the possibility that bankers' behaviour is affected by beliefs regarding other bankers' behaviour. This belief, in addition to beliefs about what one should do, guides typical norm obedience.[55] Cohn et al[4] assess expectations of banker behaviour in the coin-flipping task, by incentivising a separate group of bankers (n=142) to predict bank employees' reported winning tosses. In their study, this norm obedience manipulation did not work, and bankers primed with their professional identity did not raise expectations of bankers' dishonesty (p=0.921, rank-sum test (two sided)).

In an extension to the original study, we attempted a more direct measure of peer expectations. We asked bankers to report expectations of colleagues' average reported winning tosses in the coin-flipping task, after completing the task. We found some evidence among the ME sample, of primed bankers having marginally greater expectations of their peers reporting greater winning coin tosses ($\text{mean}_{\text{treatment}}$ = 5.6, $\text{mean}_{\text{control}}$ = 5.1, p=0.074, rank-sum test (one-sided), n=148) though not in the AP sample ($\text{mean}_{\text{treatment}}$ = 5.4, $\text{mean}_{\text{control}}$ = 5.4, p=0.684, rank-sum test (one-sided), n=620). The ME banker finding is potentially indicative of weaker honesty norms associated with banking given the positive correlation with higher winning coin tosses (spearman's $\text{rho}_{\text{two-tailed}}$ = 0.274, p< 0.01) though may also reflect a post-decision rationalisation of dishonesty or a 'false consensus effect'.[56]

c. Competitiveness

i. Compensation structures

Cohn et al[4] speculate that priming professional identity may have prompted associations with competitive remuneration schemes - schemes which are more common in core business units as opposed to those providing support services. They find some evidence of elevated cheating among core business units in their probit analysis (p=0.008, Wald test), however, they find no interaction with the saliency of professional identity (p=0.960, Wald test).

In our regression analyses, across the Asia Pacific and Middle East samples we do not find that working in a 'core unit' is associated with more dishonesty in any of our models (Extended Data Tables 1A and 1B*)*. When pooling banker data from the original and our studies, the core unit is significantly correlated with higher reports of winning coin tosses (p=0.049, p=0.048, Chi-squared test, in models b and c, respectively), however, similar to Cohn et al, no interaction with priming professional identity is found (p=0.702, p=0.689, Chi-squared test, in models b and c, respectively). We note that classifications of bankers into core and support service units are necessarily different in our studies given the different nature of services offered by commercial banks. For consistency, we strived to separate staff conducting support services from other roles (e.g. customer-facing roles).

ii. Self-reports of competitiveness

Cohn et al[4] also speculate that the culture of banking may be one in which competitiveness is considered to be a desirable behaviour. As such, priming professional identity in combination with the competitive nature of the coin-flipping task, may generate dishonesty among treated bankers. Competitiveness is measured by responses on a seven-point scale to the question: 'How important is it to you to be the best at what you do?' They find no evidence for competitiveness increasing with priming banker identity (p=0.461, rank-sum tests (one-sided), n=128) and find no correlation between competitiveness (independent of priming) and winning coin flips in their regression analysis (p=0.642, Wald test).

Among the AP bankers, higher self-reports of competitiveness were not found among those primed with their professional identity (p= 0.602, rank sum test (one-sided), n=620). In probit analysis, competitiveness was not found to be associated with a higher probability of reporting a winning coin toss (Extended Data Table 1A).

Among the ME bankers, higher self-reports of competitiveness were not found among those primed with their professional identity (p=0.874, rank sum test (one-sided), n=148). Still, those self-reporting higher levels of competitiveness, independent of priming, were more likely to report a higher number of winning coin flips, according to probit analyses (p=0.021, Wald test (Extended Data Table 1B)). No interaction was found between professional identity priming and competitiveness in probit analyses.

Pooling our banker data with Cohn et al, increased competitiveness was not found among bankers primed with their professional identity (p=0.707, rank sum test (one-sided), n=896), nor was competitiveness (independent of priming) associated with higher reports of winning coin flips (Extended Data Tables 1C and 1D). Together, these results suggest that competitiveness, as measured by self-reports, is not a channel through which priming generated increased dishonesty.

2.2. Non-banking Employees

2.2.1. Descriptive statistics

A balanced sampling technique was used to recruit a population in the Asia Pacific jurisdiction that was representative in terms of gender and age. All participants who joined the study completed screening questions aimed at ensuring that they currently resided in the desired jurisdiction and were employed either part-time or full-time.

a.

	Cohn et al			Asia Pacific		
	Control	**Treatment**	**All**	**Control**	**Treatment**	**All**
Participants (number)	66	67	133	125	117	242
Age (average years)	46.0	45.1	45.5	42.2	43.9	43.0
Males (proportion - %)	84.8	91.0	88.0	50.4	48.7	49.6
Professional Experience (average years)	14.1	15.5	14.8	11.5	14.8	13.1
University Education (proportion - %)	84.8	68.7	76.7	47.2	48.7	47.9
Relative Income (7-point scale)	5.52	5.36	5.44	3.90	4.13	4.01

	Middle East			Europe		
	Control	**Treatment**	**All**	**Control**	**Treatment**	**All**
Participants (number)	38	29	67	109	96	205
Age (average years)	38.9	42.2	40.4	39.5	40.5	40.0
Males (proportion - %)	60.5	79.3	64.2	46.8	54.2	62.0
Professional Experience (average years)	10.2	9.7	10.0	4.8	6.3	5.5
University Education (proportion - %)	97.4	93.1	95.5	90.8	87.5	89.3
Relative Income (7-point scale)	4.00	4.10	4.05	3.91	4.03	3.97

There are a few apparent differences between the samples of non-bankers. Most noticeably, Cohn et al's[4] sample had a higher representation of men relative to those in the Asia Pacific, Middle East and European samples. The original study's sample contained a higher proportion of people with a university education relative to the Asia Pacific sample, though lower than that observed in the Middle East and European samples The original study's[4] sample also has more professional experience on average than the other samples, and consistent with that, a perception of higher relative income. None of these variables - gender, university education, professional experience, relative income were found by Cohn et al[4] to affect the probability of reporting of winning coin tosses.

2.2.2. Randomization checks

Non-banking participants were randomly and evenly allocated between treatment and control conditions by the survey platform, Qualtrics. Non-parametric tests were used to assess whether proper randomisation had been achieved. Wilcoxon rank sum tests (two-sided) were used for interval variables, while Chi-squared tests (one-sided) were used for binomial variables. The estimates are broadly consistent with randomisation being achieved.

a. Asia Pacific Non-bankers

	Total Sample (n=242 participants)		**Control (n=125 participants)**		**Treatment (n=117 participants)**		
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	p-value
Age (average years)	43.0	14.0	42.2	13.7	43.9	14.2	0.399
Males (proportion - %)	49.6	0.501	50.4	0.502	48.7	0.502	0.794
Professional Experience (average years)	13.1	21.6	11.5	11.0	14.8	28.9	0.433
University Education (proportion - %)	47.9	0.501	47.2	0.501	48.7	0.502	0.814
Relative Income (7-point scale)	4.01	1.38	3.90	1.39	4.13	1.36	0.160

b. Middle Eastern Non-bankers

	Total Sample (n=67 participants)		**Control (n=38 participants)**		**Treatment (n=29 participants)**		
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	p-value
Age (average years)	40.4	10.8	38.9	10.0	42.2	11.6	0.193
Males (proportion - %)	68.7	0.467	60.5	0.495	79.3	0.412	0.101
Professional Experience (average years)	10.0	7.5	10.2	6.8	9.7	8.5	0.399
University Education (proportion - %)	95.5	0.208	97.4	0.162	93.1	0.258	0.403
Relative Income (7-point scale)	4.05	1.70	4.00	1.76	4.10	1.66	0.771

c. European Non-bankers

	Total Sample (n=205 participants)		**Control (n=109 participants)**		**Treatment (n=96 participants)**		
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	p-value
Age (average years)	40.0	11.4	39.5	10.6	40.5	12.4	0.600
Males (proportion - %)	50.2	0.501	46.8	0.501	54.2	0.501	0.292
Professional Experience (average years)	5.5	6.1	4.8	5.4	6.3	6.8	0.037*
University Education (proportion - %)	89.3	0.310	90.8	0.290	87.5	0.332	0.443
Relative Income (7-point scale)	3.97	1.32	3.91	1.30	4.03	1.33	0.400

Note that perfect randomisation was not achieved among the European sample of non-bankers with regard to professional experience. Those in the treatment condition self-reported having more years in their industry than those in the control condition. The original study found no relationship between professional experience and the likelihood of reporting a winning outcome in both the main (n=128) and follow-up (n=80) studies.

2.2.3. Priming checks

Similar to the banking professionals, we attempted to assess the effectiveness of professional identity priming among the financial services regulator samples with the use of a word quiz. We did not attempt a manipulation check of the Asia Pacific non-banker sample, in-line with Cohn et al[4], as we held the concern that the diversity of professions did not lend itself to such an effort.

The use of a quiz as manipulation check for regulators was made more challenging by absence of a previously validated task for regulators and the inability to pre-test the task for effectiveness among regulators - a seemingly more inaccessible population than bankers.

Nevertheless, financial services regulators were asked to solve six word puzzles, four of which of had financial and regulatory-themed solutions. Consistent with the bankers, comparisons of the number of professionally-themed solutions between the control and treatment group is used to assess the effectiveness of priming regulatory professional identity.

a. Middle Eastern Financial Services Regulators (Non-bankers)

The word puzzles presented to the Middle Eastern financial services professionals were a combination of two finance-themed puzzles from Cohn et al, and two regulatory-themed puzzles that we developed.[57]

_ o n e y	-> money
B _ n d	-> bond
_ u l e	-> rule
_ i n e	-> fine

The number of puzzles solved was not significantly different between the treatment and control groups (mean$_{treatment}$ = 2.2, mean$_{control}$ = 2.0, p=0.269, rank-sum test (one-sided)). This may be interpreted to suggest that the induction of professional identity failed, as those in the treatment group did not solve more puzzles. Alternatively, it may indicate that the subjects in the control condition also had a strong financial regulatory identity, independent of priming.

b. European Financial Services Regulators (Non-bankers)

In light of the apparent failure of the manipulation check with the Middle Eastern regulators we reviewed the word puzzle task. Given the low correct response rate to the 'fine' puzzle across both conditions, this puzzle was replaced with the following puzzle: _ a w (law). No other changes were made.

In this sample, we found that participants primed with their professional identity, were slightly more likely to solve word puzzles with finance and regulation-themed words (mean$_{treatment}$ = 1.9, mean$_{control}$ = 1.7, p=0.042, rank-sum test (one-sided)).

2.2.4. Treatment effects

Non-banker Reported winning outcomes from the coin-flipping task (average)	**Control Condition**	**Treatment Condition**	**p-value** (one-sided rank-sum test)
Asia Pacific (n=242)	5.48	5.85	0.114
Middle East (n=67)	5.11	5.03	0.472
Europe (n=205)	5.26	5.22	0.572

a. Asia Pacific

The mean number of reported wins out of 10 coin flips was 5.85 for the treatment group and 5.48 for the control group, recalling that this sample excluded all participants who self-reported currently working in the banking sector. According to the Wilcoxon rank-sum test (one-sided) [subsequent p-values relate to the same test, unless otherwise stated], the treatment group did not report a statistically different outcome to the control group (p=0.114). This result held when removing the 38 participants (16 in the treatment group, 22 in control) who previously (though not currently) held a role in the banking sector (p=0.293).

Of note, if we take a subset of participants (n=38) who report being bankers with previous experience in banking, a treatment effect is found (p=0.040). While this could potentially raise an interesting question regarding the sustainability of internalised norms throughout individuals' careers, our sample is insufficiently powered (n=16 in treatment, n=22 in control), to make any robust conclusions.

b. Middle East

The mean number of reported winning coin flips from 10 rounds was 5.03 for the treatment group and 5.11 for the control group. The difference between the two groups was not statistically significant (p=0.472).

c. Europe

The mean number of reported wins out of 10 coin flips was 5.22 for the treatment group and 5.26 for the control group. The difference in reported outcomes between the two groups was not statistically significantly (p=0.429).

2.2.5. Dishonesty estimates

To determine if any statistically detectable dishonesty was occurring in the aggregate level in both conditions, comparisons were made between the theoretical distribution of outcomes from a fair coin and the actual control and treatment group distributions.

a. Asia Pacific.

Participants in both the control and treatment conditions engaged in statistically detectable levels of cheating (p=0.034 and p=0.002, respectively).

b. Middle East

Relative to the fair coin distribution, neither the control group nor the treatment group reported higher levels of winning outcomes (p=0.515 and p=0.570, respectively).

c. Europe

Relative to the fair coin, neither the control group nor the treatment group reported higher levels of winning outcomes (p=0.188 and p=0.260, respectively).

2.2.6. Expectations of dishonesty

We asked the non-banking participants in the Asia Pacific their expectations of reported winning outcomes about different populations, using a between subjects design similar to Cohn et al.[4] Descriptive statistics are provided in the table below.

a.

Expectations of Reported Outcomes from Coin-flip Task (0-10 wins)		
	Mean	Standard Deviation
Medical Doctors (n=55)	6.04	1.53
Bankers (n=65)	6.34	2.04
Prison Inmates (n=64)	6.23	2.21
General Population (n=58)	5.95	1.78

No differences are found between the expectations of reported outcomes from the coin-flip task (p-value =0.559, Kruskal-Wallis chi-squared = 2.065) across the four groups. Consistent with this, undertaking pair-wise comparisons between expectations of bankers' reported winnings and those of medical doctors, prison inmates, expectations are not found to differ in any case (p=0.340, p=0.523, p=0.165, respectively; Bonferroni-corrected significance level =0.05/3=0.017). In line with the original study, the pair-wise comparisons are made using two-sided rank-sum tests. This finding contrasts with the original study, which found that expectations banker behaviour in the coin-flipping task were indistinguishable from those of prisoners, tended to be worse than the general population and were significantly worse than for medical doctors.

These differences could also stem from different sampling approaches. We used a panel provider to source a nationally representative sample (n=242) based on gender and age, whereas the original study (n=183) drew upon a convenience sample of male visitors to a Municipal Office.

We find no strong evidence for heterogeneous gender effects in our sample of Asia Pacific non-bankers. Using, two-sided rank-sum tests, we find tentative indications for women holding higher expectations of reported winning outcomes among doctors and prisoners, though these are not significant when correcting for multiple comparisons (p=0.059, p=0.033; corrected significance level = 0.05/4 =0.013). Further, there is no indication of any gender-based differences for bankers or the general population (p=0.537 and p=0.552, respectively).

2.3. Sample comparisons

2.3.1. Cross-national cultural analyses

Differences in national culture could be a factor which explains variation in the effect sizes from priming professional identity. Indeed, there is evidence from a study spanning 23 countries that differences in national norms regarding rule violations predict dishonesty.[5] As such, differences in jurisdictions and their associated national norms, may cause the 'base-line' honesty in each jurisdiction to differ. We make an assessment of that by comparing the non-banker samples and untreated bankers from the undisclosed jurisdiction of the original study, and those in the Asia Pacific and the Middle East, respectively.

In the absence of treatment effects among non-bankers, the treatment and control groups were not separated for the analyses. No differences were found between the samples of non-bankers from the original study (n=133), and those from the Asia Pacific (p=0.542, two-sided Wilcoxon rank-sum test, n=242). In the Middle East, we find that the non-bankers reported less winning outcomes (50.7%) than the non-bankers from the original study (57.8%, p=0.006, one-sided Wilcoxon rank-sum test). We are cautious about drawing conclusions from this comparison given the small (n=67), non-representative nature of the Middle Eastern study (e.g. only financial services regulatory professionals were surveyed).

Reported Outcomes from Coin-flip Task for Non-bankers (0-10 wins)		
	Mean	Standard Deviation
Cohn et al (n=133)	5.78	1.89
Asia Pacific (n=242)	5.66	1.91
Middle East (n=67)	5.07	1.75

Turning to the untreated bankers in the various jurisdictions, we also find no differences when conducting two-sided rank-sum tests and correcting the p-value for multiple comparisons between those in Cohn's et al's[4] jurisdiction with those in the Asia Pacific (p=0.395) and the Middle East (p=0.178) - where dishonesty is detected unlike in the original study.

Further, using measures of rule violations (e.g. national measures of corruption, tax evasion, and fraudulent politics) found to correlate with dishonesty,[5] we find negligible differences between an advanced Western economy (the presumed original jurisdiction, though is not possible to be confirmed for legal reasons) and our jurisdiction in the Asia Pacific, though notable differences with the Middle East. The latter finding suggests that national cultural differences may have played a role in explaining variation in results between the original jurisdiction and Middle East, as the elevated level of 'baseline' dishonesty among the non-primed bankers may have contributed to a smaller treatment effect.

Overall, these analyses suggest that national differences and honesty norms are unlikely to be responsible for the differences in findings between the AP study and that of the original study, though we cannot so confidently rule out an influence in the Middle Eastern study with the available data.

2.3.2. Banker and Non-banker analyses

We explored whether there were differences in honesty between bankers and non-bankers in each jurisdiction in order to gain an insight into the relative standing of bankers. In comparing the reported winning outcomes of Cohn et al's[4] unprimed participants, non-bankers reported more winning outcomes than bankers (p=0.002). Indeed, Cohn et al[4] found no evidence that the unprimed bankers engaged in dishonest reporting. By contrast, in the Asia Pacific samples of unprimed of bankers and non-bankers, both samples are found to engage in dishonesty when comparing their outcomes to those of a theoretical fair coin distribution (p=0.029 and p=0.021, respectively). Further, no differences are found in reported winning outcomes for unprimed bankers and non-bankers in the Asia-Pacific (53.0% vs. 54.8%, p=0.202).

While less reliable due to the small sample sizes in addition to methodological and sampling variations, in the Middle East we also no find differences between unprimed bankers and non-bankers (54.9% vs. 51.1%, p=0.109). We would note that the control group of Middle Eastern bankers tended not report winning outcomes that differed from those of a fair coin (p=0.051 - SI 2.1.4), suggesting that people with more honest tendencies may have self-selected into the banking industry in that sample, similar to the original jurisdiction.

We also explored differences between reported winning outcomes among treated bankers in the original study's jurisdiction and treated bankers the Asia Pacific and Middle Eastern jurisdictions. Again, we find no differences (p=0.206 and p=0.765 respectively, Wilcoxon rank-sum test (two-sided)). Still, we cannot eliminate the possibility and indeed, likelihood of differing national banking norms, nor indeed that of differing commercial and investment banking cultures.

2.3.3. Banker segment analyses

a. Reported Winning Outcomes for Bankers

	Total Sample			Control			Treatment		
Variable (0-10 wins)	Sample size (n)	Mean	S.D.	Sample size (n)	Mean	S.D.	Sample size (n)	Mean	S.D.
Cohn et al (main study)	128	5.48	1.81	67	5.16	1.64	61	5.82	1.93
Asia Pacific	620	5.34	1.71	334	5.30	1.67	286	5.40	1.75
Middle East	148	5.59	1.75	77	5.49	1.61	71	5.69	1.89
Commercial Bankers (AP & ME)	768	5.39	1.72	411	5.33	1.66	357	5.45	1.78

To make an assessment of heterogeneity in banking culture across segments of the industry ideally bankers are drawn from the one jurisdiction. While we are unable to do this due to difficulties in recruiting such samples and impossibility to know the original study's jurisdiction, when comparing the available data for investment and commercial bankers reported outcomes in the coin-flipping task, we do not find any statistical differences. Cohn et al's[4] primed investment bankers do no appear to cheat more than Asia Pacific commercial bankers, Middle Eastern commercial bankers, nor all commercial bankers combined (p=0.103, p=0.383, p=0.134, respectively). When undertaking the same comparisons with the control groups of bankers, again, no differences are found (p=0.802, p=0.911, p=0.847, respectively).

b. Self-Reported Materialism for Bankers

	Total Sample			Control			Treatment		
Variable (1-7 scale)	Sample size (n)	Mean	S.D.	Sample size (n)	Mean	S.D.	Sample size (n)	Mean	S.D.
Cohn et al (main study)	128	4.20	1.49	67	3.94	1.48	61	4.49	1.46
Asia Pacific	620	4.30	1.42	334	4.30	1.43	286	4.30	1.40

Middle East	148	4.75	1.78	77	4.57	1.83	71	4.94	1.71
Commercial Bankers (AP & ME)	748	4.39	1.50	411	4.35	1.52	357	4.43	1.49

Cohn et al[4] speculated “the professional identity prime may have increased dishonesty through an increase in materialistic values.” While we found no evidence for such channels, it may be due to commercial bankers being less materialistic than investment bankers. Accordingly, we compared a self-report measure thought to approximate materialism between the primed bankers from the original study containing investment bankers, and those from the Asia Pacific, Middle East and a combination of the Asia Pacific and Middle East, the latter representing all commercial bankers in our study. We do not detect elevated levels of materialism among Cohn et al’s sample relative to these three primed groups (p=0.148, p=0.955, p=0.337. respectively).

Reference to control groups, also suggests there are differences in this dimension, but not in the expected direction. Unprimed commercial bankers in the Asia Pacific, Middle East and both samples combined self-report higher levels of materialism (p=0.022, p=0.007, p=0.013). Whether the behaviours as opposed to self-reports of commercial bankers reveal greater materialism than investment bankers remains an open question.

c. Self-Reported Competitiveness for Bankers

	Total Sample			**Control**			**Treatment**		
Variable (1-7 scale)	Sample size (n)	Mean	S.D.	Sample size (n)	Mean	S.D.	Sample size (n)	Mean	S.D.
Cohn et al (main study)	128	5.61	1.08	67	5.64	1.06	61	5.59	1.12
Asia Pacific	620	6.07	1.05	334	6.10	1.01	286	6.05	1.09
Middle East	148	6.62	0.69	77	6.69	0.63	71	6.54	0.752
Commercial Bankers (AP & ME)	768	6.18	1.01	411	6.21	0.98	357	6.15	1.05

Similar to Cohn et al,[4] we do not find any evidence that self-reports of competitiveness have a relationship with reporting winning outcomes (see SI 2.1.6.c(ii)). Still, we compare the self-reported competitiveness of Cohn et al's investment bankers to our samples to explore possible differences in banking culture segments. We speculated that, for example, the greater variable pay found among investment bankers, would attract and possibly generate more competitiveness than that found in commercial bankers.

We compared the self-reports commercial bankers in the Asia Pacific, Middle East and aggregate with Cohn et al's investment bankers and found, against expectations, that primed commercial bankers report being more competitive than their investment banking counterparts ($p<0.01$, $p<0.01$, $p<0.01$). The same findings held when making the same comparisons among unprimed bankers ($p<0.01$, $p<0.01$, $p<0.01$). While commercial bankers report being more competitive, their performance in the coin-flipping task with its embedded competitive lottery, did not appear to induce greater competitive behaviour. Consequently, this finding does not assist in explaining the smaller effect size found among commercial bankers.

2.4. Experimenter Disclosure

To understand the effects of varying experimenter disclosure on the outcomes of a coin-flipping task, we experimentally manipulated such disclosures. Refer to Methods for more details.

2.4.1. Descriptive statistics

A total of 925 participants from Amazon Mechanical Turk drawn from the United States of America completed the survey, across the three conditions: deception (n=315), incomplete disclosure (n=309) and transparency (n=301). Of all the participants, 46% were male. The average age was 35.7 years, with a range of 18 to 76 years of age reported. Approximately 60% of participants reported having graduated from university, and nearly 13% of the sample graduating with a higher degree.

2.4.2. Randomization Checks

Participants were randomly and evenly allocated between deception, incomplete disclosure and transparency conditions. Non-parametric tests were used to assess whether proper randomisation had been achieved. Kruskal-Wallis Chi-squared tests were used for binary variables. The estimates are consistent with randomization being achieved.

a.

	Deception (n=315 participants)		**Incomplete Disclosure (n=309 participants)**		**Transparency (n=301 participants)**		**Kruskal-Wallis Chi-Squared Test**
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	p-value
Age (years)	35.1	11.1	36.7	11.7	35.2	11.6	0.113
Male	0.438	0.497	0.447	0.498	0.489	.501	0.411
MTurk Experience (years)	2.05	1.88	2.02	1.78	1.86	1.66	0.463
University Education	0.635	0.482	0.576	0.495	0.581	0.494	0.253

2.4.3. Treatment Effects

A Kruskal-Wallis Chi-squared test for differences across the conditions found a difference (p=0.0272). No ordering effects of cheating increasing with the transparency of experimenter disclosure were found (p=0.194, Jonckheere-Terpstra test (two-sided).

Comparisons were made between each of the experimental groups, using a Bonferroni-corrected significance level (i.e. 0.05/3 = 0.017) for Wilcoxon rank-sum tests (one-sided). No differences were found between the deception and incomplete disclosure conditions (p=0.188) - the experimental disclosures that Cohn et al[4] and we used, respectively.

The only difference found was between the incomplete disclosure and transparency conditions (p=0.006). Still the size of the difference between the average means that there would have been practically no impact on actually payoffs in the experiment. That is, the difference would translate to a negligible change in financial payoffs after accounting for rounding of reported outcomes (*see table below*).

	Deception (n=315 participants)		Incomplete Disclosure (n=309 participants)		Transparency (n=301 participants)	
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Reported Winning Outcomes (from 10 coin flips)	5.84	2.13	5.99	1.95	5.64	2.10

Of course there are limitations around generalizing this finding. This study was conducted on a sample of US Amazon Mechanical Turk workers and therefore, may not hold among bankers or other populations. It could be the case that such Amazon Mechanical Turk workers, given their experience participating in many studies and/or pressure to complete the study in a timely fashion[58], are less attentive to the disclosure purpose of an experiment. Consequently, this is an avenue for future research.

3. Supplementary Notes

3.1. Power Analysis

a. Bootstrapped samples - statistical significance

We determined the likelihood of finding a statistically significant result from rank-sum tests from simulated samples. Samples were drawn with replacement in 50 subject increments up to a sample size of 1000. The samples were drawn from Cohn et al's main study[4], where a treatment effect was detected.

An even number of bankers were drawn from the control and treatment groups, in keeping with the random allocation of the experimental design. For each sample size, 10,000 simulated samples were drawn and rank-sum tests conducted. The proportion of statistically significant outcomes from these tests was calculated, using a threshold of alpha = 0.05.

b. Bootstrapped samples - effect size

We determined the likelihood of finding the same or larger effect size as Cohn et al. The effect size was measured by the difference in means of average reported coin tosses (0.655) as per the main study the original paper.[4]

We drew samples of differing sizes with increments ranging from 50 to 1000. An even number of control and treatment subjects were drawn, in keeping with the random allocation of the experimental design. The samples were drawn with replacement from our Middle Eastern and Asia Pacific banker samples, separately. We calculated the mean reported coin tosses for the control and treatment groups and compared the difference in means to that of Cohn et al. Repeating this 10,000 times, we determined what proportion of simulated samples generated a difference in means from the control and treatment groups greater than or equal to that of the main study in Cohn et al.

c. Recruitment

In undertaking recruitment efforts, only medium to large-sized banks were approached to participate in the study, in an effort to improve the power of the experiment by obtaining a larger sample size than Cohn et al. Given the nature of field studies, it was not, however, possible to pre-specify the sample size of the study, given that participation and completion rates of the survey are unknown prior to running the study.

3.2. Funding

All studies were fully-funded by Google ATAP, via the London School of Economics and Political Science.

4. Human Subjects Approval

Approval for the study was provided by the Research Ethics Committee of the London School of Economics and Political Science (Reference Number 000582).

5. Additional References

53. Becker, G. S. (1974). Crime and Punishment: An Economic Approach. In W. M. Landes (Ed.), *Essays in the Economics of Crime and Punishment* (pp. 1-54). UMI (National Bureau of Economic Research)

54. Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self Concept Maintenance. *Journal of Marketing Research ,* **45** (6), 633-644.

55. Bicchieri, C. (2006), The Grammar of Society: The Nature and Dynamics of Social Norms, *Cambridge University Press*.

56. Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, *13*(3), 279-301.

57. Koopman, J., Howe, M., Johnson, R. E., Tan, J. A., & Chang, C. H. (2013). A framework for developing word fragment completion tasks. *Human resource management review*, *23*(3), 242-253.

58. Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior research methods*, *44*(1), 1-23.

Chapter 5: Manuscript of Project 2

Rahwan, Z., Hauser, O. P., Kochanowska, E., & Fasolo, B. (2018). High stakes: A little more cheating, a lot less charity. *Journal of Economic Behavior & Organization*, 152, 276-295. https://doi.org/10.1016/j.jebo.2018.04.021

Zusammenfassung auf Deutsch

Wir untersuchen die nachgelagerten Konsequenzen des Betrügens - und des Widerstehens der Versuchung zu betrügen - bei hohen Einsätzen auf pro-soziales Verhalten und Selbstwahrnehmung. In einer großen Online-Stichprobe replizieren wir das bahnbrechende Ergebnis, dass die Betrugsraten weitgehend unempfindlich gegenüber der Höhe des Einsatzes sind, selbst bei einer 500-fachen Erhöhung. Wir präsentieren zwei neue Ergebnisse. Erstens führte das Widerstehen der Versuchung, bei hohen Einsätzen zu schummeln, zu einem negativen moralischen Übertragungseffekt, der einen moralischen Freibrief auslöste: Teilnehmer, die unter der Bedingung hoher Einsätze dem Schummeln widerstanden, spendeten anschließend einen kleineren Teil ihres Gewinns für wohltätige Zwecke. Zweitens schätzten Teilnehmer, die maximal schummelten, ihre wahrgenommene Moral falsch ein: obwohl diese Teilnehmer dachten, dass sie sich weniger unmoralisch fühlten, wenn sie schummelten, fühlten sie sich einen Tag nach der Betrugsaufgabe unmoralischer als unmittelbar danach. Wir diskutieren die theoretischen Implikationen unserer Ergebnisse zu moralischem Abwägen und Selbstbetrug sowie die praktische Relevanz für die Gestaltung von Organisationen.

Journal of Economic Behavior and Organization 152 (2018) 276–295

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

High stakes: A little more cheating, a lot less charity

Zoe Rahwan[a,b], Oliver P. Hauser[b,c,d], Ewa Kochanowska[e], Barbara Fasolo[a,*]

[a] London School of Economics and Political Science, London WC2A 2AE, UK
[b] Harvard Kennedy School, Cambridge, MA 02138, US
[c] Harvard Business School, Boston, MA 02163, US
[d] University of Exeter Business School, Exeter EX4 4PU, UK
[e] IESE Business School, Barcelona 08034, Spain

ARTICLE INFO

Article history:
Received 4 October 2017
Revised 12 April 2018
Accepted 25 April 2018
Available online 11 June 2018

Keywords:
Cheating
Incentives
Moral licensing
Moral self-perceptions
Pro-social behaviour

ABSTRACT

We explore the downstream consequences of cheating–and resisting the temptation to cheat–at high stakes on pro-social behaviour and self-perceptions. In a large online sample, we replicate the seminal finding that cheating rates are largely insensitive to stake size, even at a 500-fold increase. We present two new findings. First, resisting the temptation to cheat at high stakes led to negative moral spill-over, triggering a moral license: participants who resisted cheating in the high stakes condition subsequently donated a smaller fraction of their earnings to charity. Second, participants who cheated maximally mispredicted their perceived morality: although such participants thought they were less prone to feeling immoral if they cheated, they ended up feeling more immoral a day after the cheating task than immediately afterwards. We discuss the theoretical implications of our findings on moral balancing and self-deception, and the practical relevance for organisational design.

1. Introduction

Corporate misconduct and unethical behaviour remain a widespread problem in organisations, ranging from large-scale fraud (e.g., Bernie Madoff, Enron) to smaller, everyday unethical behaviours (Zhang et al., 2015; Sezer et al., 2015). Organisations often rely on compensation schemes to foster better performance among employees but, at the same time, those incentives could also encourage more cheating inadvertently to meet requirements of such schemes. Recently, for example, employees at Wells Fargo who were financially incentivised for every newly opened bank account created over 2 million bank accounts illegally without customers' permission.

Because the size of incentives can affect motives and behaviour in often unexpected ways (Gneezy and List 2014; Gneezy and Rustichini 2000a,b), past research has paid particular attention to the size of incentives and its effects on immediate opportunities for cheating. A somewhat surprising, yet consistent finding is that the size of incentives for behaving unethically does not affect rates of unethical behaviour much in laboratory studies. We take this research in a new direction by looking at the downstream effects of cheating–or, resisting the temptation to cheat–on future moral behaviour. In particular, we explore the behavioural and psychological consequences of providing an immediate opportunity to cheat at varying sizes of stakes on subsequent charitable giving.

[*] Corresponding author at: London School of Economics and Political Science, London WC2A 2AE, UK.
E-mail address: b.fasolo@lse.ac.uk (B. Fasolo).

https://doi.org/10.1016/j.jebo.2018.04.021
0167-2681/© 2018 Elsevier B.V. All rights reserved.

Mazar et al. (2008) document that a four-fold increase in incentives did not change average levels of cheating in a laboratory experiment. In a recent large-scale replication, Kajackaite and Gneezy (2017) observe little change in cheating rates as incentives were increased to much higher stakes than previously studied (up to $50 in a U.S. laboratory setting) when there is some chance of cheating being detected.[1] What explains this refusal to cheat more at higher stakes? Mazar et al. (2008) hypothesise that people attempt to maintain a self-concept of an honest person. Thus, they will cheat to increase their payoff but do so only to the extent that it does not negatively affect their moral self-image.

The idea of maintaining a "moral self" has also been proposed in a separate research stream on moral balancing (Miller and Effron 2010; Zhong et al., 2009). This research suggests that people balance their good and bad behaviour over time: when people feel they have sufficiently established that they are a moral person, they become more likely to engage in immoral behaviour in the future ("moral licensing"). For example, Effron et al. (2009) demonstrate that participants who could endorse Barack Obama were more likely to feel licensed in a subsequent decision to choose a white applicant over a black applicant in a hypothetical hiring decision, having previously established their moral credentials as a non-racist moral decision-maker.

Here we combine the literature on stake size on immediate cheating opportunities with research on moral balancing to explore whether different levels of incentives affect subsequent moral behaviour. We hypothesise that, when given the opportunity to cheat, participants will resist this temptation at low and high stakes; however, resisting the temptation to cheat at high stakes will be psychologically taxing and provide participants with a plausible reason to excuse future transgressions of moral behaviour, exhibiting a moral licensing effect.

To explore the psychological costs and benefits of dynamic moral behaviour, we measure participants' self-perceptions of their morality. While cheating and donations are behavioural, incentive-compatible outcomes, self-reports are simply stated views–however, as such they inform how people (would like to) view themselves and provides an insight into the psychological processes of how cheating is dealt with. We predict that cheating a little does not affect moral self-perceptions; only maximal cheating–an unambiguous signal of immoral behaviour–negatively affects self-perceptions, an effect that persists (or even worsens) over time.

To test these predictions, we recruit a large-scale sample of participants ($N = 2015$) on Amazon Mechanical Turk (MTurk). We begin by exploring the role of stake size on dishonest behaviour using a modified version of the "mind game" (Kajackaite and Gneezy 2017; Cohn et al., 2014). We introduce significant variation in incentives, increasing stakes by up to 500 times. The maximum reward in our study is US $50 for our 10 minute online task – a significant amount for workers on MTurk whose median reservation wage has been estimated at US $1.40 per hour (Horton and Chilton, 2010).

In the second part of the study, we explore the downstream effects from honest (or dishonest) behaviour. First, we collect self-perceptions of morality using self-reports. Second, we give participants the opportunity to give to charity, allowing them to choose how much, if any, of their earnings from their first part of the study they wish to donate to a charity of their choice, a measure of whether they engaged in moral licensing by not giving to charity at all or decreasing their charitable donation. Finally, we invited participants back to the study one day later to assess their self-perceptions of their (non-)cheating behaviour the prior day.

We find that participants were not more likely to cheat when the stakes for cheating increased but with rising stakes, they subsequently donated a smaller percentage of their earnings to charity, consistent with a moral balancing account. Moreover, we find that self-perceptions vary over time: while they are stable for honest participants and those who cheat little, we find that maximal cheaters feel less moral one day after the cheating task, independent of the stake size. Surprisingly, however, self-perceptions did not meet expectations: maximal cheaters initially thought they would be less likely to feel guilty after doing wrong, when, in reality, they were the only group of participants that felt worse upon reflecting on their cheating behaviour.

1.1. Dishonesty

Research on dishonesty and unethical behaviour has attracted much attention given its apparent frequency and high costs it imposes for societies. In recent decades, much work has explored the limits of Becker (1968) utilitarian approach for understanding dishonest behaviour. Under this model, rational individuals weigh the benefits of dishonest behaviour with the chance and consequences of getting caught. Becker's model does not pay explicit regard to psychological costs of dishonest behaviour (e.g. Mazar et al., 2008, Shalvi et al., 2011, Kajackaite and Gneezy, 2017) nor the related social, organisational or political context (e.g. Gneezy, 2005, Cohn et al., 2014).

This traditional model of cheating has been challenged by findings that people are generally insensitive to payoffs for dishonesty (e.g. Mazar et al., 2008, Fischbacher and Foÿlmi-Heusi, 2013). Experimental evidence suggests that most people exhibit at least some degree of aversion to lying (Cappelen et al., 2013; Gneezy et al., 2013; Kajackaite and Gneezy, 2015). Abeler et al. (2016) show in a review of 72 cheating studies with a maximum payoff of $50 that individuals are largely insensitive to external stakes for cheating. Moreover, participants displayed almost no increase in cheating, even when stakes

[1] In another condition, Kajackaite and Gneezy (2017) make detection of cheating almost impossible: the combination of no detection and high stakes does lead to an increase in cheating behaviour, suggesting that an important boundary condition to the literature reviewed here is that participants will cheat more when it is completely impossible to detect their cheating behaviour at the individual level. We discuss the implications of these boundary conditions on our results at the end of the paper.

rose to as much as $110 (Hilbig and Thielmann 2017). In fact, individuals frequently engage in low-level cheating and commonly resist maximal cheating, leaving on average three-quarters of the maximum possible payoff on the table (Abeler et al., 2016).

Mazar et al. (2008) posit a theory of self-concept maintenance to explain why people do cheat but much less than Becker's theory would predict. They argue that individuals attempt to balance the tension between a desire to enrich themselves materially from dishonest behaviour with maintaining a favourable moral self-appraisal. It is therefore possible for individuals to consider themselves honest but nonetheless cheat "only a little bit" because partial dishonesty does not threaten their positive self-image. Gneezy et al. (2017) offer an alternative explanation of the observation that people cheat a little bit: they find that cheating is more common when participants are unobserved rather than observed, and that the frequency of partial lying increases when the maximal outcome is less likely *ex ante*. They argue that social identity–that is, the socially constructed part of an individual's self-concept–shapes what we consider to be appropriate and, consequently, accounts for non-maximal cheating behaviours.

Still, there are occasions where individuals do respond to rewards. Gneezy et al. (2017) and Abeler et al. (2016) find more cheating when participants are unobserved rather than observed. Further, many experiments have relied on the "cheating game" (e.g. Mazar et al., 2008, Fischbacher and FoÏlmi-Heusi, 2013) which makes individual detection unlikely but possible. However, Kajackaite and Gneezy (2017) argue that eliminating all concerns for detection can in fact drive up cheating rates: indeed, when detection is made impossible through a modified version of the "mind game" (Jiang 2013; Shalvi and De Dreu 2014; Potters and Stoop 2016), Kajackaite and Gneezy find that higher stakes (ranging between $1 and $50 in a laboratory setting) do lead to an increase in cheating rates. Conversely, it is important to note that, unless detection by the experimenter is made completely impossible, social identity concerns are likely going to affect the intrinsically-perceived cost of lying (Gneezy et al., 2017): participants who want to maintain a positive self-image are likely also concerned about their appearance towards a potential observer, including the experimenter, and as such increased cheating at a higher stake size might signal a negative image towards themselves and the potential observer.

In sum, social identity is an important element of self-image preservation: people's beliefs about being a moral or immoral person are shaped by whether others view them as moral or not, and these notions can become internalised. As such, people's self-perception of morality is, at least in part, socially constructed but also deeply internalised about what is right and wrong. This implies that it is unlikely that participants will cheat more at higher stake sizes relative to lower stake sizes because by doing so they would appear to be "a liar" both to themselves and others; meanwhile, cheating "just a little bit" (at smaller and higher stakes) may not have that same stigma and participants are thus likely to do so.

Taken together, these published findings suggest that people engage in at least some levels of cheating regardless of stake size, but that cheating rates do not increase with stake size.

H1. *Participants in all conditions will cheat at least to some extent.*

H2. *Participants will cheat only a little more as the stakes increase.*

1.2. Moral consistency and moral balancing

After engaging in a moral or immoral behaviour, two potential follow-on behaviours can occur. On the one hand, moral behaviour may be followed by more moral behaviour while immoral behaviour is followed by more immoral behaviour–a "consistency" account of morality (Zhong et al., 2009). Research on consistency has found that people behave consistently when they think their prior action was justified (Jordan and Monin 2008), want to reassure (and update their own beliefs about) themselves based on their prior actions (Ariely and Norton 2008), or want to avoid looking like a hypocrite (Cialdini et al., 1995). For example, the well-known Foot-in-the-Door paradigm shows that larger requests for one's time or commitment are more likely to be accepted when it was preceded by a smaller request (Freedman and Fraser 1966). This is in part driven by the fact that people derive some understanding of their own preferences based on what they observe themselves doing (Bem, 1972; Ariely and Norton, 2008).

Moral consistency usually occurs when the behaviour can be linked to an individual's identity concern. Specifically, Conway and Peetz (2012) demonstrate that, when people reflect what their actions imply for their self-image of a moral decision-maker, they are more likely to act morally, as compared to thinking about the specific actions in isolation. Furthermore, they also find that temporal distance from a particular action can affect the need to compensate behaviour or be consistent with it: behaviour that is morally good and further in the past is seen to reflect on one's identity and leads to consistency, but when a good action was taken only recently, moral compensation seems acceptable (Conway and Peetz 2012). Conversely, moral consistency can also be primed by products which are morally questionable and when the user is aware of the inauthenticity of the product: Gino et al. (2010) find that participants who knew that they were wearing counterfeit sunglasses behaved less morally than those who knew they wore authentic sunglasses.

In sum, moral consistency is a possible consequence after taking an earlier moral action. Thus, in our setting, if consistency is expected, then those who cheat the most are also less likely to choose to donate and make smaller donations.

H3. *Participants who cheat more are (i) less likely to give to charity and (ii) give a lower percentage of their earnings to charity.*

On the other hand, moral behaviour at an earlier point in time could give license to less moral behaviour later, especially when the moral behaviour just recently preceded another moral action (Conway and Peetz 2012). Research on moral balancing (Miller and Effron, 2010; Zhong et al., 2009; Merritt et al., 2012) suggests that moral decisions are not viewed in isolation but usually within the context of previous and future decisions that help establish and maintain one's moral self-image. If the goal is to maintain a moral self-image overall, one need not act perfectly moral all the time: a moral decision today might invite an immoral decision tomorrow (moral licensing) and conversely an immoral decision might be followed by a moral one to make up for bad behaviour (moral cleansing).[2]

For example, Monin and Miller (2001) show that participants that have the chance to disagree with blatantly sexist statements are more likely to later hire a man for a stereotypically male job in a hypothetical hiring decision. Likewise, Effron et al. (2009) demonstrate that endorsing Obama can lead to racial discrimination in a later decision. Even hypothetical, counterfactual or potential thinking in the future can license other moral decisions (Effron et al., 2012; Effron et al., 2013; Gneezy et al., 2012). Conversely, knowing that an opportunity to "cleanse" immoral behaviour will exist in the future leads to more unethical behaviour early on (referred to as "conscience accounting," see Gneezy et al., 2014). Cojoc and Stojan (2014) also find that people behave more immorally if they are aware of an opportunity to behave morally in the near future. Further, they find that individuals who are aware of an opportunity to donate in a subsequent part of the experiment tend to give less than those who are not informed of that opportunity, independent of whether they behaved honestly or dishonestly in a cheating task.[3]

While the existence of moral licensing and cleansing across domains is well-established (for an introduction and overview, see Miller and Effron 2010), we know relatively little about the cognitive processes and psychological costs that moral balancing requires. Here we offer one account for when and how participants engage in moral balancing when stakes matter.[4] Resisting the temptation to cheat is cognitively taxing, as it requires self-control not to give in to cheating to one's benefit (e.g. Gino et al., 2011, Greene and Paxton, 2009). We argue that the temptation to cheat increases for at least some people when the stakes for cheating rise, making it harder for them to resist the temptation. As a consequence, we predict that people who feel particularly virtuous and moral about their past behaviour–especially when the stakes are high—are more likely to feel justified later to behave more immorally. Indeed, the literature on moral licensing argues that people have an incentive and a tendency to view their benefits and costs in an asymmetric fashion (Miller and Effron, 2010). They are more likely to play up the good deeds they have done, even if just in their own minds, but downplay any negative behaviour. Due to this self-serving asymmetry, it is very likely that the higher the cheating benefits the participant resists, the higher their perception of their morality, leading to more licensing behaviour.

Assuming that participants exhibit some form of licensing behaviour after doing a good deed, past research suggests that there exist at least two possible explanations for how people might engage in moral licensing (Miller and Effron, 2010). The "moral credits" model proposes that people keep an internal "moral balance sheet:" good deeds add to the moral balance while bad deeds are like debits on the balance sheet. This model suggests that people use their moral balance to "purchase" themselves the right to a moral transgression in the future (Hollander, 1958; Merritt et al., 2012). That is, even a blatant and unambiguous moral transgression is acceptable to a moral-credits decision-maker assuming that he or she has accumulated the moral credits previously to make up for it.

In contrast, the "moral credentials" model (Miller and Effron, 2010; Effron et al., 2009) proposes that participants who behave morally in one decision are more likely to construe later moral transgressions as ambiguous and not immoral, having previously established that they are a moral decision-maker. Miller and Effron succinctly describe what a decision-maker might ask themselves: " 'Can I say or do this without signaling something morally discrediting about myself?' " (Miller and Effron, 2010: 119). Therefore a moral-credentials decision-maker would not view an ambiguous moral transgression as immoral because they have established credentials that show otherwise.

Both the "moral credits" and "moral credentials" models predict licensing after behaving morally but they differ in the pathway to getting there. In our experiment, participants make two subsequent decisions regarding giving to a charitable cause, each operating through a different pathway. Specifically, we ask participants if they would like to donate at all to charity and if so, how much of their earnings they would like to donate.

Participants choosing whether and how much to donate might display the two distinct pathways of moral licensing. The first—consistent with "moral credits" —implies that participants who have resisted cheating are *less likely* to donate anything to charity as stake size increases. Choosing not to give to charity is an unambiguous signal the decision-maker is sending after having previously behaved morally in the high stakes condition. A smaller fraction of donors as stake size increases is evidence for the "moral credits" account.

[2] An open question which is beyond the scope of this paper is to theorise about the dynamics and moderators that explain when moral consistency and moral balancing occur; for reviews and perspectives, see Dolan and Galizzi (2015) and Truelove et al. (2014).

[3] These findings may raise concerns in our study that the decision whether to cheat or not could be affected by knowledge of another moral task in the future. In our experimental design, however, participants did not know that they would later be asked to make a donation decision, so they could not have anticipated or adjusted their behaviour accordingly.

[4] In an earlier conception of this research, we hypothesised that our experimental setup would trigger moral cleansing (see link and abstract in the pre-registration in the online appendix): if participants cheat more in the coin-flipping task as the stakes increase, they might be more likely to donate a larger percentage to charity later, to "morally cleanse" for cheating at a higher rate. However, only later did we realise that the repeated mind-game was unlikely to lead to increased cheating in the coin-flipping task. Based on the fact that participants did *not* cheat in the first task, we then looked at the flip-side of moral balancing (comprising of moral cleansing and moral licensing) and predicted that this self-controlled behaviour at higher stakes would lead to more moral licensing when choosing how much to donate.

The second is consistent with the "moral credential" explanation: As stake size increases, participants give a *smaller proportion* of their earnings to charity. Donating a low percentage of a high earning is an ambiguous signal of licensing: when the stakes are high, participants who give a smaller percentage still donate a large absolute amount of their earnings to charity: as such they can construe their actions not to be immoral, thus behaving less morally without taking a hit to their moral self-image. As stake size increases, a lower percentage of earnings donated is evidence for the "moral credentials" account.

In sum, the two pathways to moral licensing make the following distinct but not mutually exclusive predictions about subsequent donation behaviour:

H4a. *("Moral Credits" Model): Participants in the high stakes condition will be less likely to donate to a charity.*

H4b. *("Moral Credentials" Model): Participants in the high stakes condition will give a smaller percentage of their earnings to charity.*

1.3. Self-reported morality

Moral balancing is simultaneously an internal, cognitive process as well as an outward projection: participants continuously evaluate their actions against a backdrop of previous behaviours and situations and justify the morality of their actions to themselves and others (Zhong et al., 2009; Jordan et al., 2011; Shalvi et al., 2011, 2012). Thus, we are not just interested in measuring behaviour such as cheating or donations, as perceived by outside spectators, but also in participants' self-reports of their own morality – that is, how participants perceive their (im)moral behaviour, and how those self-reports might predict their future moral behaviour.

Participant morality relating to a situation may be assessed by implicit or explicit measures. Implicit measures include solving word fragments with terms which may or may not relate to morality. For example, Gino et al. (2011) used word fragment puzzles which could be solved with morality-related terms such as 'moral', 'virtue' or unrelated terms such as 'mural' or 'tissue.' Implicit elicitation of morality is useful in circumstances, where, for example, concerns of social desirability prevail, or moral awareness may not be activated (Bazerman and Banaji, 2004; Rudman, 2004).

In contrast, explicit self-reports of morality involve asking a participant how they feel or anticipate feeling in regard to a moral dilemma (e.g., having the opportunity to cheat on a coin-flipping task). Self-reports are helpful in understanding how a decision-maker views themselves, and portrays themselves to others, after acting (un)ethically (Rudman, 2004). Effron et al. (2015) ask participants, for example, how guilty, unethical, dishonest they feel after such a task. Similarly, Zhong and Liljenquist (2006) elicit self-reports of moral emotions after engaging in a physical cleansing task. Gino and Desai (2012) use 'moral purity' to assess how 'innocent' and 'morally pure' participants felt.

We predicted that the higher stakes of cheating in some conditions would lead to more temptation to cheat, which requires effort to resist but which also provides participants with an opportunity to claim and leverage virtuosity. We therefore expect that self-reports of morality would provide a clue to understanding the presence of moral licensing. Specifically, resisting temptation to cheat when the stakes are high might lead some participants to feel highly moral, which they leverage to justify their subsequent immoral behaviour. We consider two outcomes to measure (*i*) an effect on self-reported morality on average as stakes increase and (*ii*) a tendency to leverage perceptions of high morality to excuse less moral behaviour subsequently.

H5a. *Self-reported morality increases with higher temptation to cheat.*

H5b. *Those who perceive themselves as highly moral in the high stakes condition give a smaller percentage of their earnings to charity.*

1.4. Reflecting on morality

We have argued that resisting the temptation to behave dishonestly could lead to unexpected downstream consequences in a subsequent moral decision and that self-perceptions of morality can be used to justify immoral behaviour. But what happens after the 'heat' of the cheating moment has passed?

Empirical research on the management of a positive self-view after unethical behaviour (Chugh and Kern, 2016), and in particular reflections on unethical behaviour to manage a positive self-view, has received little scholarly attention. Only recently, Kouchaki and Gino (2016) proposed a pathway through which people might deal with their unethical past. The authors give participants the opportunity to cheat a little. Several days later, they invite the same participants back to the study and ask them to remember the details of the cheating (versus non-cheating) task. They find that participants obfuscate their past unethical behaviour by forgetting the details of the cheating task they had to complete. They engage in "unethical amnesia" – the subconscious and purposeful forgetting of details of their past unethical acts (Kouchaki and Gino, 2016).

Here we investigate reflections on feelings of morality—not the factual details of the task—one day after the completion of a cheating task. Kouchaki and Gino's research suggests that those who cheat a little engage in obfuscation of their actions and do not report feeling morally different than those who behaved completely honestly.

However, we propose that obfuscation of the memory of feeling immoral does not apply to "obvious cheaters" who are in clear violation of the rules: when behaviour is unambiguously immoral (such that there is no "moral wiggle room" (e.g.,

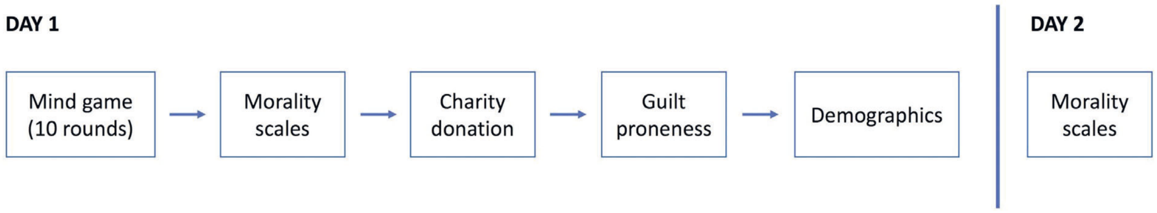


Fig. 1. Illustration of each stage of the two-day study.
Note. Participants played the repeated mind game for 10 rounds, followed by scales that elicited their self-reported morality. Then participants were asked if they wanted to give some of their earnings to a charity and, if so, what percentage of their earnings. Participants then filled out a number of exploratory items (including the 'negative behavioural evaluation' subscale of the Guilt and Shame Proneness (GASP) scales which we refer to as "guilt proneness" scale) and demographics. All participants received an invitation the next day to participate in a short follow-up survey that included the same morality scales from the day before.

Dana et al., 2007) to pretend otherwise), feelings of morality will—perhaps surprisingly—be lower than those who cheat only a little and, importantly, become worse over time. Specifically, we propose:

H6a. *Participants who cheat a little feel morally (i) no different than those who behave completely honestly and (ii) no different upon reflection a day later.*

H6b. *Participants who cheat maximally, such that their behaviour is unambiguously immoral, will feel (i) morally worse than those who do not cheat or cheat only a little and (ii) feel even worse when reflecting on their morality a day later.*

Finally, as the stakes increase, we would expect this effect to increase, such that the overall effect of maximal cheaters who feel guiltier as time passes arises primarily from the high stakes conditions, where maximal cheating yields the greatest payoff.

H6c. *The effect of maximal cheaters feeling worse upon reflection is strongest in the highest stakes condition.*

2. Experimental design

2.1. Experimental design overview

Our experiment was conducted on Amazon Mechanical Turk (MTurk)[5] over two days in January 2017 (see Fig. 1). On the first day, we recruited a total of $N = 2{,}015$ participants across four conditions. Participants received $1.00 for participating in the study and were told that they had the possibility to earn a bonus payment based on the decisions they made during the study. Since the maximum possible bonus varied by condition, the exact amount of potential bonus earnings was not disclosed at the time of recruitment.

In all conditions, participants first read the instructions for a coin-flipping task. All participants engaged in 10 rounds of the coin-flipping task in which they had the opportunity to cheat to earn an additional bonus payoff. We randomly allocated participants to one of four conditions, which varied the maximum potential payoff that they could earn from the coin-flipping task ($0.10, $0.50, $5.00, or $50.00). All payoff-relevant decisions were incentive-compatible.

After the coin-flipping task, participants self-reported their feelings of morality and indicated what percentage (if any) of their earnings they want to give to a charity of their choice. The participants finished the study by responding to a few vignettes regarding guilt proneness and filling out demographic information.

On the second day, we invited all participants back to complete a short survey. A total of $N = 1{,}413$ participants (70%) returned for the second survey.[6] In the second survey, participants reported whether they won in the lottery draw and again self-reported their feelings of morality from the coin-flipping task completed a day earlier.

[5] Amazon Mechnical Turk is an online labour market which has been used extensively for economic research; however, it does not come without limitations and potential selection effects. We refer interested readers to Clifford et al. (2015), Horton et al. (2011) and Landers and Behrend (2015) for a discussion of the limitations and opportunities of the MTurk platform for research purposes.

[6] We tested for differential uptake of the second survey. We did not find any correlation between uptake and cheating behaviour in the coin-flipping task, self-reported morality, donation likelihood or donation level using logit models (using logit to predict uptake of the second survey by each of the independent variables: all *ps* > 0.05). There were, however, some differences in uptake across conditions (Kruskal-Wallis=8.23, *p*=0.042) though trend effects from higher stakes were only marginally significant (Jonckheere-Terpstra=778240, *p*=0.072). In conducting pairwise comparisons between the conditions, we used bonferroni-corrected *p*-values. The only significant difference was between the $5 and $50 maximal payoff conditions (Wilcoxon=116780, *p*=0.03, bonferroni-corrected) such that more participants in the $50 condition completed the second survey. All other post-hoc comparisons were not significant (*ps* > 0.1).

2.2. Measuring dishonesty

Our experiment involves participants playing a variant on the traditional cheating game introduced by Jiang (2013). Tasks that involve flipping a coin or rolling a die in private are unobtrusive measures of honesty (e.g. Bucciol and Piovesan, 2011). Traditional coin-flipping tasks reward the participant based on a pre-determined outcome, as specified by the experimenter (e.g., the participant gets paid if the coin comes up heads). The participant flips the coin in private and then reports the outcome. While this task has been shown to lead to moderate rates of cheating, it has been argued that it does not conceal a participant's dishonesty completely because the participant might think that the experimenter would be able to detect their dishonesty if they are able to check the actual outcome of the coin flip.

The 'mind game' variant of the traditional cheating game enables a more robust concealment of dishonesty from the participant's perspective. Specifically, in our design, participants are asked to (*i*) think of an outcome from a coin toss and remember it, (*ii*) toss a coin in private and report the outcome, and (*iii*) report whether the actual outcome matched the outcome they had thought of. In this setup, the experimenter is unable to verify, on an individual level, whether the reported outcome matched what the participant thought of previously. Participants play 10 rounds of the same coin-flipping mind game. Because the participant knows that the experimenter cannot access his or her thoughts, the participant can be assured that he or she will not be detected. Indeed, Kajackaite and Gneezy (2017) have demonstrated that cheating rates in a (single-round) mind game increase at higher stakes compared to the equivalent traditional cheating game. While the experimenter cannot identify on an individual level which participant cheated, it can be determined whether cheating occurred at a group level based on the theoretically expected probabilities of matches occurring in the group. The average cheating rates can thus be compared across conditions to detect whether more or less cheating occurred as incentives changed.

Participants' potential earnings depended on the number of reported coin flips matching their imagined outcome. We varied the incentive to cheat across four conditions: in condition 1, participants earned US $0.01 per matching coin flip outcome; in condition 2, US $0.05 per match; in condition 3, US $0.50 per match; and in condition 4, US $5.00 per match. Across all 10 rounds of the game, participants could therefore earn a maximum of US $0.10, US $0.50, US $5.00 and US $50.00 in conditions 1 through 4, respectively. The condition with a maximum potential earning of US $50.00, which we will refer to as "high stakes" condition, represents significant stakes for crowd-sourcing platforms like MTurk where reservation wages of US $1.4 dollars per hour have been measured (Horton and Chilton, 2010).

A competitive lottery mechanism was used to award earnings. The lottery approach was used to manage the costs of running the experiment and has been found not to distort true preferences (Starmer and Sugden, 1991). In particular, following the procedure by Cohn et al. (2014), participants were told that their number of coin-flip matches would be compared to another randomly selected participant; if the participant reported an equal or higher number of coin flips to the comparator participant, they were entered in the lottery draw. In the lottery, one in five subjects are randomly selected as winners and received their earnings as a cash bonus payment on the second day of the experiment.

2.3. Measuring self-reported morality

We used self-reports of morality to measure the costs of cheating, or resisting the temptation to cheat, on both day 1 and 2 of the experiment. After completing the mind game coin-flipping task, we asked subjects to complete self-reports on how moral, virtuous, dishonest and unethical they felt on a 5-point Likert scale (where 1 = not at all, slightly, somewhat, very much, 5 = extremely). The adjectives were presented in a randomised order, and a combined 'morality' index for analysis was created.

The same four morality items were assessed again in the follow-up study. A day later after the coin-flipping experiment, we invited participants to answer the same questions about their moral feelings in relation to the coin-flipping task they completed a day earlier.

2.4. Donation behaviour

After completing the coin-flipping task and reporting their morality on day 1 of the experiment, participants were asked if they would like to donate any of the potential earnings from the coin-flipping task to a charity of their choice. Participants were told they could choose from a list of six popular US charities.[7] If participants chose to answer the first question—"would you like to donate some or all of the bonus to a charity?" —with "yes," they were presented with two follow-up questions: (*ii*) "Which charity would you like to donate some or all of any bonus awarded?" and (*iii*) "What percentage (%) of any bonus awarded would you like to donate to your chosen charity?" To enable comparability across conditions, we asked what proportion of their potential earnings they would like to donate (0–100%). The slider scale was anchored at 50% and increments of 25% were available.[8]

[7] The charities to choose from were the Red Cross, St Jude Children's Research Hospital, Salvation Army, United Nations Children's Fund (UNICEF), American Society for the Prevention of Cruelty to Animals (ASPCA), Habitat for Humanity. We selected these charities based on their popularity and brand recognition.

[8] The survey software used, Qualtrics, requires that the slider scale have an anchor.

2.5. Control and exploratory variables

At the end of the survey, participants filled out a short survey to control for individual variation and to allow for exploratory analysis. The items included the 'negative behavioural evaluation' subscale of the Guilt and Shame Proneness (GASP) scales, materialism, competitiveness (Cohen et al., 2011). In addition, we collected information on demographics, past donation behaviour and frequency, and MTurk experience both in terms of years and past participation in coin-flipping tasks.

2.6. Pre-test experiments and power analysis

Before conducting the main experiment we ran a pre-test experiment ($N = 180$). The pre-test was conducted for three reasons. First, we aimed to pre-test and correlate all stated measures in a non-overlapping sample of MTurk participants with actual cheating behaviour. Specifically, we explored implicit and explicit approaches to measuring morality: after the "mind game," participants were randomly assigned to either solving word fragments with terms relating to cheating costs (e.g. b_d -> bad, _ _ eater -> cheater)[9] or self-reported measures of feeling virtuous, moral, dishonest and unethical (as described in Section 2.3), respectively. We adopted aspects of the self-report scales from Effron et al. (2015), which measured guilt and virtue after a cheating task.

We were interested whether actual cheating behaviour over 10 rounds of the same "mind game" as in the main experiment (as described in Section 2.2) affected morality, as measured implicitly by the word fragment puzzle or explicitly through self-reports of moral emotions (or both). Unlike the main experiment, however, there was no randomisation of stake size; for all participants, the total stake size across the ten rounds was fixed at $0.50 (i.e., the equivalent of condition 2). We found significant positive correlations between the matches reported in a cheating task and self-reports of feeling dishonest and unethical (rho = 0.35, $p = 0.003$). However, we did not find a significant relationship between matches reported and the other outcomes, including morality-related words solved for in the word fragment task (all $ps > 0.05$). We thus decided to retain the self-reported measures and exclude the word fragment task in the main experiment.

Second, we aimed to determine the minimum sample size needed to detect small changes in our main outcome variables in the main study. For conservative measure, we focused on a subset of participants in the second pre-test experiment who completed the moral self-reports and reported not knowing the purpose of the coin-flipping task. We simulated 1,000 random samples for differently-sized groups, and determined the proportion of samples for each group that found a significant positive correlation between the number of coin flips reported as matched and negative moral affect using a one-sided Spearman test at the 5% significance level. Using this procedure, we determined that a sample size of at least 460 participants per condition was required to achieve 80% power in the main experiment. We decided to round up to 500 participants per condition to further increase power.

Finally, we were interested in testing whether including the self-reported morality scales before the donation decision would affect this subsequent moral decision. We found that neither measure of morality (self-reports and solving word fragments) significantly affected the subsequent likelihood of engaging in a charitable donation nor the proportion of earnings to be donated (both $ps > 0.05$). We thus decided to keep the morality measures between the cheating and donation tasks for the main experiment.

3. Results

3.1. Dishonesty: little cheating for big money

We begin by looking at the effects of stake size on cheating behaviour in the coin-flipping task. First, we look at the average number of times participants in our experiment reported a 'match.' Given sufficient number of observations, the expected theoretical value from ten coin-flips per participant is 5 matches if we assume all participants act entirely honestly. However, the actual number of reported matches in our dataset is 6.28, above the expected value. Using a two-sided one-sample t-test, we reject the null hypothesis that the expected value equals the actual reported average number of matches per participant, $t(2{,}014) = 31.94$, $p < 0.001$. Furthermore, we conduct a nonparametric comparison between the theoretical distribution expected under full honesty (i.e., the theoretical distribution of a fair coin) and the observed empirical distribution of matches reported: we detect a significant difference between the distributions ($Z = 2{,}817{,}200$, $p < 0.001$), suggesting the presence of cheating. Thus, in line with our prediction in H1, we find significant levels of low-level cheating in our sample.

Second, we compare cheating behaviour across the four conditions with varying incentives to cheat per coin-flip. We find that increasing stake size leads to significantly more cheating but the differences between smaller and larger stakes are economically small (Fig. 2**A**); the distributions of matches reported across 10 rounds are qualitatively similar (Fig. 2**B–E**). Formally, we find significant differences between the distributions of matches across conditions using a Kruskal–Wallis test ($H(3) = 17.884$, $p < 0.001$). We also conduct a Jonckheere–Terpstra test to compare the ordering of the number of matches reported across the four conditions and find a significant effect of increasing stake size on the level of cheating, ($T_{JT} = 810{,}720$,

[9] Gino et al. (2011) use this method to measure access to ethics-related words.

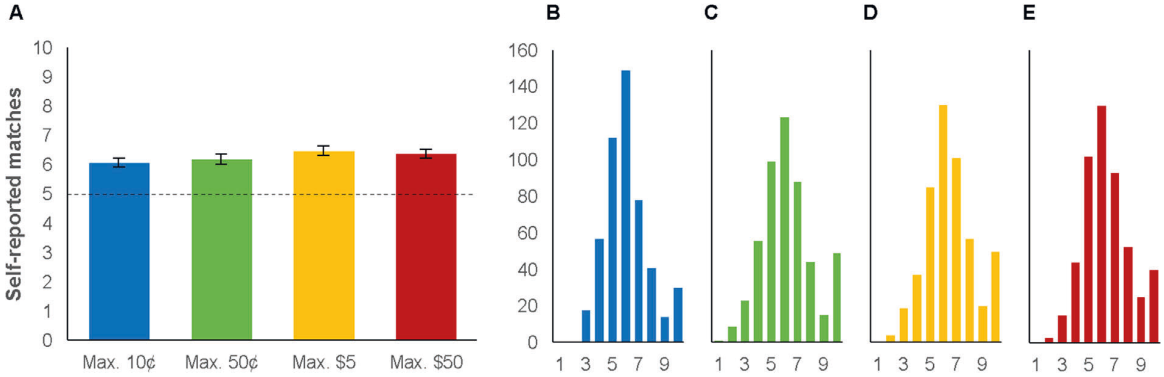


Fig. 2. Dishonest behaviour changed little as incentives increased.
Note. (A) As incentives increase, participants report significantly more matches in the 10 rounds of the coin flipping task. However, the effect size is significant but small: the average number of matches increases by less than 5% as incentives multiply by a factor of 500. *(B-E)* The distributions of matches reported across 10 rounds by increasing stake sizes (maximum payoffs: (B) $0.10, (C) $0.50, (D) $5.00, or (E) $50.00) look qualitatively similar. .

$p < 0.001$). Bonferroni-corrected post-hoc pairwise comparisons with Wilcoxon tests show that there were significant differences between the $0.10 and $5.00 condition ($W = 107{,}900$, $p < 0.001$), $0.10 and $50.00 condition ($W = 113{,}640$, $p = 0.005$), and the $0.50 and $5.00 condition ($W = 115{,}140$, $p = 0.007$). However, there were no significant differences between the two low-stakes condition ($0.10 and $0.50, $p > 0.10$) and the two high-stakes condition ($5.00 and $50.00, $p > 0.10$).

These results suggest that a 500-fold increase in the size of stake to cheat increases cheating only very little. To quantify the extent of cheating by stake size, we predict average matches with a continuous variable measuring the increases in stake size. Fitting a linear regression using a continuous variable with the four stake sizes in cents (1, 5, 50, and 500) reveals only a weak, marginally significant relationship with cheating behaviour (coeff = 0.0003, $p = 0.101$; see Online Appendix Table A1). A log-transformation of the stake size, instead, is a better predictor in the same model (coeff = 0.055, $p = 0.001$), suggesting that each 10-fold increase in stake size leads to reporting only a 0.05 higher number of average matches reported (Table A1).

Taken together, these results support our second hypothesis (H2) that, while cheating does significantly differ across conditions, the effect sizes are economically small, such that participants are generally insensitive to stake size.

3.2. Moral consistency and balancing

3.2.1. Smaller percentage donated at higher stakes

We next focus on understanding the downstream consequences of cheating, or resisting the temptation to cheat, on future moral behaviour. We first hypothesised that those who cheat more are also generally less likely to give to charity and give less to it. Conversely, however, for some participants resisting the temptation to cheat will be psychologically taxing the larger the stakes are, and consequently, we predicted that higher stakes can trigger moral licensing in a subsequent moral decision.

To examine the potential consistency and balancing response, we look at the effect of stake size on the probability to make a donation and the donation amount. Using these two outcomes, we are able to distinguish between the "moral credits" —measured by the likelihood to donate—and "moral credentials"—measured by the percentage of earnings donated—explanations for licensing.

First, we observe that the probability to make a donation does not differ by condition (using logit regression predicting making a donation of any amount by condition dummies: all $ps > 0.1$, Table A2) (Fig. 3**A**). Furthermore, we also do not find that those who are more likely to have cheated (i.e., higher number of matches reported) are any more (or less) likely in the high stakes conditions to donate (using donation likelihood predicted by interaction of condition and total number of matches: all $ps > 0.1$), though total number of matches independently predicts lower likelihood to donate across conditions (without interaction: coeff = –0.150, $p < 0.001$; with interaction terms: coeff = –0.179, $p = 0.003$). This suggests that (*i*) those who cheat more are less likely to give to charity, lending support that at least some of our participants are engaging in moral consistency (H3) and (*ii*) there is no evidence of blatant moral licensing at higher stakes in the form of lower likelihood to donate, suggesting that these results do not provide support for the "moral credits" model (H4a).

Conversely, the "moral credentials" model suggests that stake size might affect donation levels, a more ambiguous behaviour to license. The donation level is the percentage of earnings from the coin-flipping task that a participant chooses to donate, provided they selected to make a donation in the previous question. We find that donation levels do vary significantly with condition (Fig. 3**B**): relative to the $0.10 condition, participants in the $5.00 condition (coeff = −13.403, $p < 0.001$, Table 1) and $50.00 condition (coeff = −19.850, $p < 0.001$) donate significantly less to charity but not in the $0.50 condition ($p > 0.1$). All other pairwise differences between conditions are significant ($ps < 0.05$).

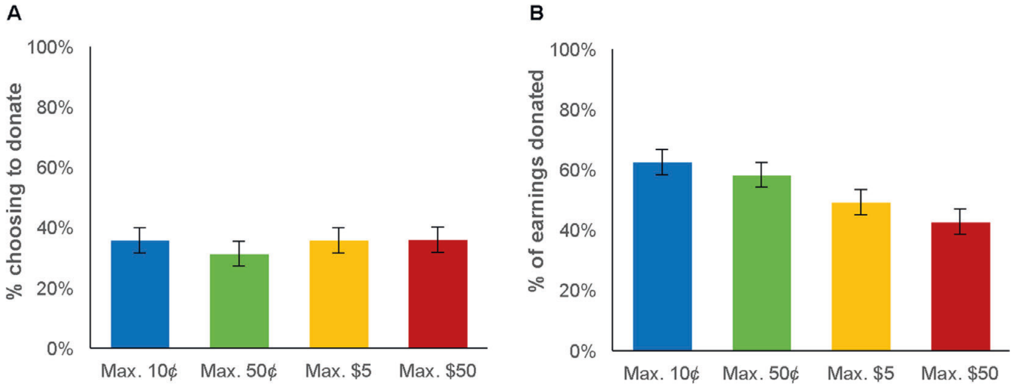


Fig. 3. Higher stakes did not affect the likelihood to engage in charitable behaviour but decreased the percentage of earnings donated.
Note. (A) The likelihood of making a donation does not change with stake size. *(B)* However, participants who choose to donate are affected by the size of their potential earnings: as stake increases donors choose to give away a smaller share of their earnings to a charity of their choice. This effect is driven by self-reported morality: participants who report feeling more moral in the high-stakes condition donate less to charity.

Table 1
Donation levels—the percentage of earnings donated—are lower in higher-stakes conditions.

Variables	(1)	(2)	(3)
Condition: Max. $0.10	(Omitted)	(Omitted)	(Omitted)
Condition: Max. $0.50	−4.236	−3.810	−4.266
	(3.296)	(3.222)	(12.186)
Condition: Max. $5.00	−13.403***	−12.116***	−19.994
	(3.132)	(3.096)	(12.346)
Condition: Max. $50.00	−19.850***	−18.931***	−28.730*
	(2.861)	(2.825)	(12.236)
# Matches reported		−3.300***	−4.082**
		(0.665)	(1.486)
Max. $0.50 X # Matches reported			0.094
			(1.974)
Max. $5.00 X # Matches reported			1.327
			(1.967)
Max. $50.00 X # Matches reported			1.654
			(1.961)
Constant	62.570***	81.634***	86.148***
	(2.305)	(4.531)	(9.084)
Observations	700	700	700
R-squared	0.072	0.105	0.107

Note. Linear regression predicting percentage of earnings donated by condition dummies and self-reported number of matches in the coin-flipping task (for those participants who decided to donate). Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

These results remain qualitatively robust to including the number of matches as a covariate: participants in both the $5.00 and $50.00 conditions donate significantly lower percentages of their earnings to charity ($ps < 0.001$). Conversely, the coefficient on the number of matches reported is significantly negative, giving support to the notion that some participants were exhibiting moral consistency (H3; coeff = −3.300, $p < 0.001$). Finally, we consider interactions with the number of matches reported, and find that the treatment effect is not dependent on the likelihood to cheat ($p > 0.1$ for all interaction terms). That is, those who are more likely to have cheated do not donate any more or less in the high stakes conditions. Taken together, these results provide support for the predictions of the “moral credentials” model (H4b).

3.2.2. Leveraging high morality to justify lower donations

To understand how participants cognitively process giving less to charity, we explore participants’ self-reported morality. Morality was assessed immediately after, and in reference to, the coin-flipping task. We first look at average levels of morality across conditions, followed by individual-level correlations between moral self-reports and donation behaviour.

Across all conditions, we do not find significant differences in the average levels of morality reported (Table 2; using linear regression predicting self-reported morality by condition dummies; followed by pairwise comparisons of coefficients: all $ps > 0.1$). There are also no interactions with the behaviour in the coin-flipping task (using linear regression with condition dummies interacted with number of total matches reported: all $ps > 0.1$), though the total number of matches reported in

Table 2
Self-reported morality does not change with stake size, but is predicted by the number of matches reported in the coin-flipping task.

Variables	(1)	(2)	(3)
Condition: Max. $0.10	(Omitted)	(Omitted)	(Omitted)
Condition: Max. $0.50	−0.071	−0.062	0.093
	(0.046)	(0.045)	(0.163)
Condition: Max. $5.00	0.028	0.059	0.230
	(0.046)	(0.046)	(0.171)
Condition: Max. $50.00	−0.016	0.006	0.259
	(0.046)	(0.045)	(0.172)
# Matches reported		-0.076^{***}	-0.052^{**}
		(0.009)	(0.019)
Max. $0.50 X # Matches reported			−0.025
			(0.026)
Max. $5.00 X # Matches reported			−0.028
			(0.026)
Max. $50.00 X # Matches reported			−0.041
			(0.027)
Constant	4.178^{***}	4.642^{***}	4.493^{***}
	(0.033)	(0.063)	(0.123)
Observations	2,015	2,015	2,015
R-squared	0.002	0.037	0.038

Note. Linear regression predicting self-reports of morality by condition dummies and self-reported number of matches in the coin-flipping task. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

the coin-flipping task does negatively predict morality (without interaction terms: coeff = −0.076, $p < 0.001$; with interactions: coeff of single effect = −0.052, $p = 0.008$), suggesting that participants' self-perceptions of their morality, on average, correlate with actual cheating behaviour in the coin-flipping task. These results thus do not provide evidence that participants on average inflate their morality when resisting the temptation of greater stakes (H5a).

However, the aggregate view of morality on the condition-level might obfuscate individual-level cognitive processes. We argued that resisting the temptation to cheat in the high stakes condition can be used by participants to justify their licensing behaviour (i.e., donating a smaller percentage of their earnings). That is, while participants may, on average, not believe they are more moral in the high stakes condition, some participants could construe their resistance to not give in to cheating when the stakes are high as a justification for deserving more for themselves in a future moral decision. If so, we would expect a negative correlation between participants' self-reported morality and the percentage they donate to charity in the high stakes condition, but not in other conditions.

This is exactly what we find: donors in the $50.00 condition who feel *more* moral donate *less* to charity (using linear regression predicting donation level in the $50.00 condition by morality: coeff = −6.425, $p = 0.009$, Table 3), while the morality of participants in lower-stakes condition does not predict donation levels ($p > 0.1$, except in the $0.10 condition where, in fact, donors give marginally *more* to charity the higher the self-reported morality, $p = 0.114$, suggesting that at the lowest stakes this relationship might point towards moral consistency rather than licensing). In addition, when all interaction terms between conditions and morality are included in the regression, the interaction between morality and the high stakes condition is significant (predicting donation level by condition and morality: coeff of interaction term of $50.00 condition = −11.839, $p = 0.007$; all other interaction terms, $ps > 0.1$). These results are further robust to including the number of matches reported.

In sum, while there is no evidence for inflation of morality in the high stakes condition generally, we found support for the hypothesis (H5b) that participants in the high stakes condition who claim they are particularly moral donate a smaller percentage of their earnings to charity.

3.3. Moral self-perceptions

3.3.1. Reflecting on morality makes maximal cheaters feel morally worse

Finally, we turn to investigating self-perceptions of one's morality. We begin by comparing self-reported morality immediately after the cheating decision and one day afterwards. Overall we find that participants report lower levels of morality one day after ($M = 4.123$, *s.d.* = 0.766) than immediately after completing the cheating task ($M = 4.172$, s.d. = 0.727), $t(1412) = 3.239$, $p = 0.001$.

However, specifically, we predicted that those who cheat a little would not feel different than those who behaved honestly and neither of those groups would feel different a day later. Conversely, those who cheated a lot would feel less moral in general and more so one day later.

To test these predictions, we group participants by the number of matches they reported during the coin-flipping task. We take as baseline the group of participants who reported between 0 and 5 matches: in this range participants "almost

Table 3
In the highest stake condition, self-reporting feeling more moral *reduces* the percentage of earnings that donors are willing to give to charity.

Variables	(1) Max. $0.10	(2) Max. $0.50	(3) Max. $5	(4) Max. $50	(5) All	(6) Interaction	(7) Interaction
Morality index	5.414	−0.629	−1.753	−6.425**	−0.851	5.414	4.335
	(3.408)	(3.152)	(3.695)	(2.439)	(1.581)	(3.107)	(3.058)
Condition: Max. $0.50					−4.420	21.843	27.471
					(3.079)	(18.434)	(18.134)
Condition: Max. $5.00					−13.396***	17.704	17.799
					(2.964)	(20.952)	(20.574)
Condition: Max. $50.00					−19.899***	31.138	29.294
					(2.958)	(18.943)	(18.604)
Max. $0.50 X Morality						−6.043	−7.358
						(4.291)	(4.221)
Max. $5.00 X Morality						−7.168	−6.882
						(4.777)	(4.691)
Max. $50.00 X Morality						−11.839**	−11.202**
						(4.342)	(4.265)
# Matches reported							−3.406***
							(0.659)
Constant	39.082**	60.925***	56.786***	70.220***	66.264***	39.082**	63.441***
	(14.959)	(13.207)	(16.198)	(10.570)	(7.174)	(13.640)	(14.199)
Observations	179	159	180	182	700	700	700
R-squared	0.014	0.000	0.001	0.037	0.073	0.083	0.117

Note. Linear regression predicting percentage of earnings donated by self-reported morality (separately by condition, Cols. 1–4) and condition dummies (Cols. 5–6). The number of self-reported matches in the coin-flipping task is controlled for in Col. 7. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

certainly" behaved honestly – i.e., they are statistically likely to have reported the true outcome of the coin flip.[10] We created three additional groups based on the number of matches reported, in increasing likelihood of cheating across the ten round of the coin-flipping task: group 2 contains all participants who reported 6 or 7 matches ("likely honest"), group 3 is made up of participants who reported 8 or 9 matches ("likely dishonest"), and group 4 consists of participants who reported 10 matches. We have referred to the latter as "maximal cheaters" above, given that they have "almost certainly" behaved dishonestly in at least some round of the coin-flipping task.

Consistent with our predictions (H6a), we find no large differences in morality between groups 1, 2, and 3 immediately after the cheating task (using linear regression predicting morality by group dummies: group 1 vs. 2: $p > 0.1$; group 1 vs. 3 and group 2 vs. 3: $p = 0.071$ and $p = 0.075$, respectively, suggesting a marginal decrease in morality as the likelihood of that the participants cheated increases; Table A3). Furthermore, participants in these three groups did not report a change in their morality one day later (using t-tests for each group, testing morality ratings immediately after the task with morality ratings one day later, all $ps > 0.1$).

In contrast, maximal cheaters reported significantly lower morality ratings than any other group (using linear regression predicting morality by group dummies: group 1 vs. 4: coeff = −0.693, $p < 0.001$; group 2 vs. 4: coeff = −0.689, $p < 0.001$; group 3 vs. 4: coeff = −0.597, $p < 0.001$, Table A3).[11] Moreover, those participants also reported feeling less moral than a day earlier when reflecting back on the coin-flipping task, $t(117) = 3.897$, $p < 0.001$. In fact, in support of our predictions (H6b), the group of maximal cheaters was the only group that reported reduced feelings of morality on the second day (Fig. 3**A**; using linear regression predicting difference in self-reported morality between day 1 and 2 by group dummies, see Table A4).

What role did stake size play for self-perceptions? While we expected self-reported morality to be especially low among maximal cheaters in the high-stakes condition, we do not find strong evidence for this prediction: when the difference in self-reported morality immediately after the coin-flipping task and one day later is regressed on group dummies independently for each condition, we find that this difference measure for maximal cheaters is significantly lower, as expected, in the higher-stakes conditions ($5.00 stakes: coeff = −0.370, $p = 0.026$; $50.00 stakes: coeff = −0.350, $p = 0.031$) than the $0.10 baseline, whereas there is no significant difference between the lower stakes conditions ($0.10 stakes: coeff = −0.222, $p > 0.5$; $0.50 stakes: coeff = −0.100, $p > 0.1$). While this analysis might at first suggest that stake size do play a role, we do not find significant differences when all interaction terms are included in the full regression (all $ps > 0.1$, Table A4); nor when we compare the full interaction of group and stake size on self-reported morality on the first day (all $ps > 0.1$, Table A3).

[10] In theory, participants could of course have lied and misreported the outcome of any of the single round of the game, which we cannot detect given that guessing the correct outcome of the coin-flips 50% of the time is the most likely outcome. For the purposes of the discussion here, we will refer to this group as "almost certainly honest."

[11] Results are qualitatively similar when we control for general guilt proneness in these regressions.

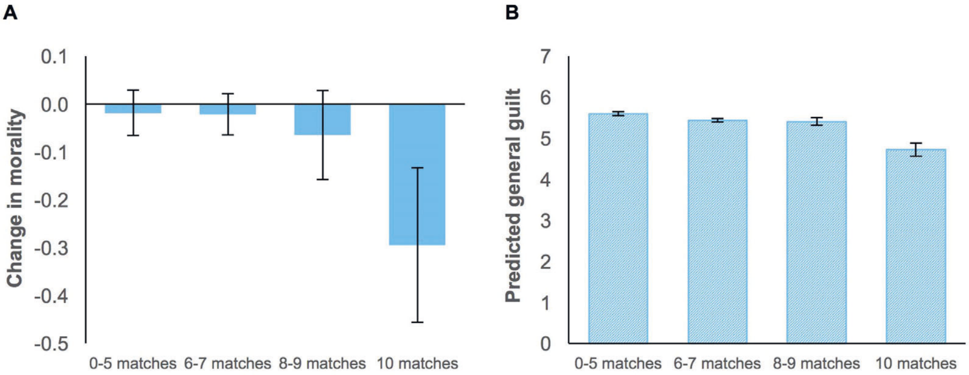


Fig. 4. Participants who cheated maximally mispredicted their perceived morality.
Note. (A) Participants who are most likely dishonest (reporting 10 matches out of 10 coin flips) report feeling less moral one day after the coin flipping task. In contrast, participants who do not cheat (0–5 matches), cheat only a little (6–7 matches), or cheat moderately (8–9 matches) report similar moral feelings the next day. *(B)* However, participants who cheat the most are bad at predicting their own feelings in the future: the most dishonest participants say they would feel less guilty after committing wrongdoing than all other participants, but their predictions do not match actual changes in reduced self-reported morality.

We conclude that there is only limited support for H6c that stake size has an effect on self-perceptions; instead, maximal cheaters, regardless of stake size, feel worse than everyone else immediately after engaging in cheating behaviour, and worse as time passes.

3.3.2. Exploratory analysis: misprediction among maximal cheaters

Finally, we turn to an additional analysis using variables which we included for exploratory purposes. While we did not *a priori* hypothesise the relationships reported below based on past literature and theory, we believe they show promising insights for future research.

Based on the fact that maximal cheaters were more likely than any other group of participants to feel worse one day later (Section 3.3.1), we became interested in understanding whether participants were able to foresee this negative change in self-perceived morality. To find out, we looked at the 'negative behavioural evaluation' subscale of the Guilt and Shame Proneness (GASP) scale. This scale captures the tendency that people would feel uncomfortable and guilty after committing an unethical act, with higher scores indicating feeling more remorse after behaving dishonestly. We refer to this scale as "guilt proneness."

Participants in general expressed that they would feel relatively uncomfortable if they committed an unethical act ($M = 5.430$, $s.d. = 1.354$). However, there exists considerable variation across participants based on their behaviour: participants who reported a higher number of matches in coin-flipping task expressed lower guilt proneness (using linear regression predicting the GASP guilt measure by total number of matches: coeff $= -0.115$, $p < 0.001$).

Perhaps surprisingly, participants who cheated maximally expressed the lowest guilt proneness compared to any group based on matches reported (linear regression predicting guilt proneness by group dummies, followed by pairwise comparisons: all $ps < 0.05$; Table A5 and Fig. 4**B**). However, as reported in the previous section, maximal cheaters were also the only group that did, in fact, show a negative change in feelings of morality one day after the task.

Taken together, these results suggest that maximal cheaters were not very good at predicting their own moral feelings: although they claimed that they would generally not feel guilty after behaving unethically, they were the only group of participants that showed a *decrease* in self-reported morality one day after committing an unethical act.

4. Discussion

We have shown that cheating only occurred at low levels, even when the stakes were extremely high for the online participant sample in this study. Yet, resisting the temptation to cheat at high stakes did have downstream effects on other moral behaviour, such as donating to charity. Specifically, we observed that participants gave a smaller fraction of their earnings to charity as the stake size increased. This suggests behaviour consistent with moral licensing: participants who refrained from cheating at higher stakes seem to have subsequently licensed themselves to donate less to charity, thereby "balancing" their moral behaviour over time. Indeed, we find that donors in the high-stakes condition who reported greater feelings of morality gave a smaller fraction of their earnings to charity. Finally, we observed a drop in self-reported feelings of morality one day after the task among maximal cheaters—but no other group of participants. This is an effect that the maximal cheaters did not appear to foresee, as they believed they were the least guilt prone after cheating.

The insensitivity to stake size we observed in our study is consistent with past work (Mazar et al., 2008; Abeler et al., 2016). In our setting, the maximum potential payoff in the lowest-stakes condition was $0.10 and in the highest-stakes condition $50.00 for a participant who was willing to cheat maximally across all rounds in the game. This marks a 500-fold increase in incentives. Given the profile of Amazon Mechanical Turk workers whose median reservation wages has been estimated as low as $1.4 per hour (Horton and Chilton, 2010), the highest payoff amount in the high-stakes condition was likely an attractive incentive for many participants. Yet, a 10-fold increase in stake size resulted only in a 0.05 increase in reported coin matches. This finding is also echoed by a recent meta-analysis: Abeler et al. (2016) re-analysed the full datasets from 46 studies across 43 countries, finding almost no effect of higher stakes across a large range of stake sizes.

However, past work has also found circumstances when cheating is sensitive to incentives. Kajackaite and Gneezy (2017) demonstrate that, when individual-level detection by the experimenter is made completely impossible, participants do cheat at higher rates when stake size increases. While we employ a similar method to Kajackaite and Gneezy (2017) and Jiang (2013), the sense of not being able to be detected is likely somewhat diminished. Participants in our experiment played several rounds of the mind game, increasing the ability of an experimenter to infer cheating behaviour at an individual level. Still, while the experimenter cannot be sure if someone cheated, participants might be less willing to engage in blatant cheating. Consequently, we observe less cheating in the multi-round mind game than past work where detection was impossible in a single-shot game.

When the stakes in the cheating task in our experiment increased, participants subsequently gave a smaller fraction of their earnings to charity. Moral licensing theory proposes that people can engage in dishonest and selfish behaviour without incurring a cost to their moral self (Miller and Effron, 2010). Our results fit a "moral credentials" explanation (Effron et al., 2009): participants leveraged their past moral credentials (i.e., resisting the temptation to cheat) to justify keeping more of their high earnings from themselves, while at the same time not seeming immoral. Merritt et al. (2010) argue that moral credentials work because the licensed action is ambiguous–immoral behaviour can be "reframed" to still be construed as moral, both to the decision-maker and others. Although participants were not aware of other conditions, participants in the higher-stakes conditions likely had an opportunity to reframe their charitable contributions in light of the fact that the amount they gave to charity (relative to typical charitable donations on Amazon Mechanical Turk) could be construed as quite high. That is, although they gave a smaller fraction of their earnings to charity, they nonetheless gave a larger absolute amount to charity than those in the smaller-stakes conditions – arguably a generous gift. Whether or not this is interpreted as morally permissible might depend on the point of view: Hauser et al. (2016) show, for example, that most participants in a group believe that group members with larger earnings ought to contribute the same (or a higher) percentage of their income to a public good, while participants with high endowments themselves think that giving a larger absolute—but not relative—amount suffices.

Overall, participants viewed themselves quite differently depending on their cheating behaviour. While a little cheating did not impact self-view of morality, participants who cheated maximally, independent of stake size, felt significantly less moral after the task. First, participants who cheat a little but do not feel immoral might engage in a mild form of self-deception. Von Hippel and Trivers (2011) argue that people engage in self-deception, often to further their own goals without paying the cost of feeling immoral – a process that Batson and Thompson (2001) call "moral hypocrisy." Self-deception may be aided by moral disengagement and motivated forgetting (Kouchaki and Gino, 2016; Shu et al., 2011; Bandura, 1999). Yet, maximal cheaters in our experiment did not attempt to deceive themselves (or others) about the morality of their actions. But, while maximal cheaters acknowledged their immoral behaviour in the moment, they seem to have misjudged the cost on their moral self-view in the future: they were the only group of participants who felt less moral a day later and, at the same time, they believed they would not feel guilty after engaging in immoral behaviour. Chance et al. (2011) demonstrate that participants consistently mispredict self-deceptive behaviour, even when it comes at a personal cost. Our results extend this line of reasoning to self-perceptions of ethical decisions more generally: even when participants believe they are fine with behaving unethically today, their view of their own actions suffers at a later date. However, we did not explicitly measure prediction of future moral self-perceptions, a task which we encourage for future research.

Our experiment is of course not without limitations. For example, the variation in donation levels we observe may in part be due to a "wealth effect," as the differences in any small or large amount earned from the cheating task could potentially affect subsequent behaviour. However, previous studies have shown that donation decisions similar to the one we employed are largely unaffected by endowment or stake size. Most relevant to our investigation, Raihani et al. (2013) study large stake size variation in the dictator game and find no wealth effects in the same Amazon Mechanical Turk population we use here. Similarly, findings from variation in endowments in an ultimatum game by Andersen et al. (2011) suggest that variation in windfall stake size would likely not explain more than 30% of the effect. While we cannot rule out wealth effects in our setting (or that our results may not be affected by extremely large stake sizes, effectively placing a boundary condition on the proposed behaviours), it is unlikely that they play a significant role in our study based on these prior findings.

Our design and findings have the following important practical implications for organisations. First, while offering a high-stakes opportunity to cheat does not translate into meaningfully higher cheating rates, organisations should carefully examine the surrounding, temporal decision context: if employees and managers feel especially morally virtuous after a high-stakes moral decision, they might engage in less moral behaviour subsequently. As such, strategies that simply deter dishonesty without eliminating temptation (e.g. by increasing monitoring or punishment, e.g. Kettle et al., 2017) may be insufficient, as they might inadvertently give a moral license. This can consequently reduce other pro-social behaviour (e.g. costly contribution to team projects, or resources directed at corporate social responsibility). Second, understanding these "spill-over" effects has implications for the design of organisational decision-making: since managers and corporate boards often face high-stakes decisions, they may be especially likely to "balance" their moral decisions – thus, a prudent intervention would assign multiple high-stakes decisions to different decision-makers. Third, when unethical behaviour coincides with a self-serving payoff, managers may subconsciously engage in some form of self-deception or downplay the likelihood that they would regret their choice. Future research could investigate whether prompting decision-makers with an intervention that informs them of the “spill-over costs” (i.e., that they might feel worse about their decision tomorrow) or that reminds them of their “moral identity” (e.g., linking it to their consistently pro-social behaviour in the past via “identity nudges,” see Kessler and Milkman, 2016) prompts momentary reflection and more ethical decision-making.

Acknowledgements

We are grateful to a grant from Google ATAP that funded this research.

Appendix

Table A1, Table A2, Table A3, Table A4, Table A5, Table A6, Table A7, Table A8, Table A9, Table A10,

Table A1
Linear regression using stake size (Cols. 1–2) and the natural logarithm of stake size (Cols. 3–4) as predictors of the average number of matches reported.

Variables	(1)	(2)	(3)	(4)
Stake size	0.000	0.000		
	(0.000)	(0.000)		
ln(stake size)			0.055^{**}	0.054^{**}
			(0.017)	(0.018)
Male		−0.169		−0.168
		(0.085)		(0.084)
MTurk experience (years)		0.036		0.035
		(0.026)		(0.026)
Familiarity with coin flips		−0.070		−0.074
		(0.089)		(0.089)
Materialism index		0.018		0.017
		(0.029)		(0.028)
Social esteem index		−0.034		−0.034
		(0.027)		(0.025)
Competitiveness index		−0.020		−0.029
		(0.031)		(0.030)
Altruism index		0.046		0.045
		(0.028)		(0.028)
Donation frequency dummies		Yes		Yes
Age group dummies		Yes		Yes
Education dummies		Yes		Yes
U.S. region dummies		Yes		Yes
Constant	6.236^{***}	6.962^{***}	6.118^{***}	6.881^{***}
	(0.048)	(0.668)	(0.064)	(0.668)
Observations	2,015	2,015	2,015	2,015
R-squared	0.001	0.031	0.005	0.035

Note. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A2
The propensity to donate to a charity does not vary by condition.

VARIABLES	(1)	(2)	(3)
Condition: Max. $0.10	(Omitted)	(Omitted)	(Omitted)
Condition: Max. $0.50	−0.199	−0.189	−0.580
	(0.134)	(0.135)	(0.488)
Condition: Max. $5.00	−0.001	0.059	−0.092
	(0.132)	(0.133)	(0.505)
Condition: Max. $50.00	0.010	0.054	−0.027
	(0.132)	(0.133)	(0.509)
# Matches reported		-0.150^{***}	-0.179^{**}
		(0.027)	(0.060)

(*continued on next page*)

Table A2 (*continued*)

VARIABLES	(1)	(2)	(3)
Max. \$0.50 X # Matches reported			0.065
			(0.078)
Max. \$5.00 X # Matches reported			0.026
			(0.079)
Max. \$50.00 X # Matches reported	–	–	0.014
	–	–	(0.081)
Constant	−0.584***	0.320	0.489
	(0.093)	(0.187)	(0.366)
Observations	2015	2015	2015

Note. Logit regression predicting choosing to donate some of one's earnings to a charity. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A3
Linear regression of self-reported morality immediately after the coin-flipping task predicted by group (based on number of matches reported) and stake size condition.

Variables	(1) Max. \$0.10	(2) Max. \$0.50	(3) Max. \$5	(4) Max. \$50	(5) All	(6) All
Group 1	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)
Group 2	−0.044	−0.041	−0.016	0.071	−0.003	−0.044
	(0.068)	(0.075)	(0.070)	(0.075)	(0.036)	(0.070)
Group 3	0.029	−0.050	−0.147	−0.190	−0.096	0.029
	(0.106)	(0.111)	(0.093)	(0.100)	(0.051)	(0.108)
Group 4	−0.588***	−0.748***	−0.607***	−0.832***	−0.693***	−0.588***
	(0.136)	(0.119)	(0.108)	(0.129)	(0.061)	(0.139)
Max. \$0.10	–	–	–	–	(Omitted)	(Omitted)
Max. \$0.50	–	–	–	–	–	−0.028
	–	–	–	–	–	(0.073)
Max. \$5.00	–	–	–	–	–	0.066
	–	–	–	–	–	(0.078)
Max. \$50.00	–	–	–	–	–	−0.004
	–	–	–	–	–	(0.076)
Group 2 X \$0.50	–	–	–	–	–	0.003
	–	–	–	–	–	(0.099)
Group 2 X \$5.00	–	–	–	–	–	0.028
	–	–	–	–	–	(0.102)
Group 2 X \$50.00	–	–	–	–	–	0.115
	–	–	–	–	–	(0.101)
Group 3 X \$0.50	–	–	–	–	–	−0.079
	–	–	–	–	–	(0.151)
Group 3 X \$5.00	–	–	–	–	–	−0.176
	–	–	–	–	–	(0.147)
Group 3 X \$50.00	–	–	–	–	–	−0.219
	–	–	–	–	–	(0.146)
Group 4 X \$0.50	–	–	–	–	–	−0.160
	–	–	–	–	–	(0.179)
Group 4 X \$5.00	–	–	–	–	–	−0.018
	–	–	–	–	–	(0.181)
Group 4 X \$50.00	–	–	–	–	–	−0.243
	–	–	–	–	–	(0.187)
Constant	4.230***	4.202***	4.297***	4.226***	4.235***	4.230***
	(0.050)	(0.054)	(0.055)	(0.057)	(0.027)	(0.052)
Observations	500	507	503	505	2,015	2,015
R-squared	0.039	0.077	0.069	0.100	0.068	0.075

Note. Groups 1, 2, 3, and 4 are made up of participants who reported 0–5 matches, 6–7 matches, 8–9 matches, and 10 matches, respectively. We refer to Group 1 as "mostly honest" while Group 4 are the "maximal cheaters." Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A4
Linear regression of difference between self-reported morality immediately after the coin-flipping task and one day later predicted by group (based on number of matches reported) and stake size condition.

VARIABLES	(1) Max. $0.10	(2) Max. $0.50	(3) Max. $5	(4) Max. $50	(5) All	(6) All
Group 1	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)
Group 2	−0.031	−0.059	0.054	0.035	−0.003	−0.031
	(0.062)	(0.062)	(0.080)	(0.073)	(0.034)	(0.068)
Group 3	−0.049	0.007	−0.106	−0.001	−0.047	−0.049
	(0.091)	(0.096)	(0.104)	(0.095)	(0.048)	(0.100)
Group 4	−0.222	−0.100	−0.370**	−0.350**	−0.276***	−0.222
	(0.124)	(0.106)	(0.122)	(0.117)	(0.058)	(0.136)
Max. $0.10	–	–	–	–	(Omitted)	(Omitted)
Max. $0.50	–	–	–	–	–	0.041
	–	–	–	–	–	(0.069)
Max. $5.00	–	–	–	–	–	−0.034
	–	–	–	–	–	(0.076)
Max. $50.00	–	–	–	–	–	−0.032
	–	–	–	–	–	(0.072)
Group 2 X $0.50	–	–	–	–	–	−0.028
	–	–	–	–	–	(0.095)
Group 2 X $5.00	–	–	–	–	–	0.085
	–	–	–	–	–	(0.100)
Group 2 X $50.00	–	–	–	–	–	0.065
	–	–	–	–	–	(0.096)
Group 3 X $0.50	–	–	–	–	–	0.055
	–	–	–	–	–	(0.143)
Group 3 X $5.00	–	–	–	–	–	−0.058
	–	–	–	–	–	(0.139)
Group 3 X $50.00	–	–	–	–	–	0.047
	–	–	–	–	–	(0.134)
Group 4 X $0.50	–	–	–	–	–	0.122
	–	–	–	–	–	(0.177)
Group 4 X $5.00	–	–	–	–	–	−0.148
	–	–	–	–	–	(0.177)
Group 4 X $50.00	–	–	–	–	–	−0.128
	–	–	–	–	–	(0.175)
Constant	−0.015	0.025	−0.049	−0.047	−0.018	−0.015
	(0.045)	(0.045)	(0.062)	(0.056)	(0.026)	(0.050)
Observations	345	356	335	377	1,413	1,413
R-squared	0.009	0.005	0.042	0.030	0.018	0.026

Note. Groups 1, 2, 3, and 4 are made up of participants who reported 0–5 matches, 6–7 matches, 8–9 matches, and 10 matches, respectively. We refer to Group 1 as "mostly honest" while Group 4 are the "maximal cheaters." Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A5
Linear regression predicting guilt proneness by group (based on number of matches reported).

Variables	(1) All	(2) Max. 10c	(3) Max. 50c	(4) Max. $5	(5) Max. $50
Group 1	(Omitted)	(Omitted)	(Omitted)	(Omitted)	(Omitted)
Group 2	−0.161*	−0.291*	−0.062	−0.141	−0.138
	(0.068)	(0.128)	(0.139)	(0.145)	(0.133)
Group 3	−0.191*	−0.334	−0.297	−0.065	−0.139
	(0.096)	(0.199)	(0.206)	(0.193)	(0.178)
Group 4	−0.878***	−0.686**	−0.884***	−0.946***	−0.900***
	(0.115)	(0.255)	(0.222)	(0.224)	(0.229)
Constant	5.600***	5.602***	5.568***	5.526***	5.700***
	(0.051)	(0.095)	(0.101)	(0.113)	(0.101)
Observations	2,015	500	507	503	505
R-squared	0.028	0.020	0.034	0.037	0.030

Note. Groups 1, 2, 3, and 4 are made up of participants who reported 0–5 matches, 6–7 matches, 8–9 matches, and 10 matches, respectively. We refer to Group 1 as "mostly honest" while Group 4 are the "maximal cheaters." Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A6
Average reported coin toss matches by stake size condition.

Condition	Average reported matches
Max. $0.10	6.08
Max. $0.50	6.19
Max. $5.00	6.48
Max. $50.00	6.37
Overall	6.28
Observations	2015

Table A7
Average donation rates and average donation proportion by stake size condition.

Condition	Average donation rates	Average donation proportion
Max. $0.10	35.80%	62.57%
Max. $0.50	31.36%	58.33%
Max. $5.00	35.79%	49.17%
Max. $50.00	36.04%	42.72%
Overall	34.74%	53.00%
Observations	2015	700

Note. Average donation proportion only for those participants who decided to donate.

Table A8
Self-reported morality immediately after and one day after the task, by stake size condition.

CONDITION	Immediately after	One day after
Max. $0.10	4.204	4.157
Max. $0.50	4.101	4.094
Max. $5.00	4.235	4.156
Max. $50.00	4.155	4.091
Overall	4.172	4.123
Observations	1413	1413

Table A9
Self-reported morality immediately after and one day after the task, by group (based on number of matches reported).

CONDITION	Immediately after	One day after
Group 1	4.242	4.224
Group 2	4.237	4.216
Group 3	4.189	4.124
Group 4	3.521	3.227
Overall	4.172	4.123
Observations	1413	1413

Note. Groups 1, 2, 3, and 4 are made up of participants who reported 0–5 matches, 6–7 matches, 8–9 matches, and 10 matches, respectively. We refer to Group 1 as “mostly honest” while Group 4 are the “maximal cheaters.”

Table A10
Guilt proneness by group (based on number of matches reported).

Condition	Guilt proneness
Group 1	5.600
Group 2	5.439
Group 3	5.409
Group 4	4.722
Overall	5.430
Observations	2015

Note. Groups 1, 2, 3, and 4 are made up of participants who reported 0–5 matches, 6–7 matches, 8–9 matches, and 10 matches, respectively. We refer to Group 1 as “mostly honest” while Group 4 are the “maximal cheaters.”

References

Abeler, J., Nosenzo, D., Raymond, C., 2016. Preferences for Truth-Telling, Center for Economic Studies and Ifo Institute (CESifo). Working Paper, No. 6087, pp. 1–118.
Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., List, J.A., 2011. Stakes matter in ultimatum games. Am. Econ. Rev. 101 (7), 3427–3439.
Ariely, D., Norton, M.I., 2008. How actions create-not just reveal-preferences. Trends Cognit. Sci. 12 (1), 13–16.
Bandura, A., 1999. Moral disengagement in the perpetration of inhumanities. Person. Soc. Psychol. Rev. 3 (3), 193–209.
Batson, C.D., Thompson, E.R., 2001. Why don't moral people act morally? Motivational considerations. Curr. Direct. Psychol. Sci. 10 (2), 54–57.
Bazerman, M.H., Banaji, M.R., 2004. The social psychology of ordinary ethical failures. Soc. Just. Res. 17 (2), 111–115.
Becker, G.S., 1968. Crime and punishment: an economic approach. In: The Economic Dimensions of Crime. Palgrave Macmillan, London, pp. 13–68.
Bem, D.J., 1972. Self-perception theory. In: Advances in Experimental Social Psychology, 6. Academic Press, pp. 1–62.
Bucciol, A., Piovesan, M., 2011. Luck or cheating? A field experiment on honesty with children. J. Econ. Psychol. 32 (1), 73–78.
Cappelen, A.W., Sørensen, E.Ø., Tungodden, B., 2013. When do we lie? J. Econ. Behav. Org. 93, 258–265.
Chance, Z., Norton, M.I., Gino, F., Ariely, D., 2011. Temporal view of the costs and benefits of self-deception. Proc. Natl. Acad. Sci. 108 (Supplement 3), 15655–15659.
Chugh, D., Kern, M.C., 2016. A dynamic and cyclical model of bounded ethicality. Res. Org. Behav. 36, 85–100.
Clifford, S., Jewell, R.M., Waggoner, P.D., 2015. Are samples drawn from Mechanical Turk valid for research on political ideology? Res. Polit. 2 (4), 1–9.
Cialdini, R.B., Trost, M.R., Newsom, J.T., 1995. Preference for consistency: the development of a valid measure and the discovery of surprising behavioral implications. J. Person. Soc. Psychol. 69 (2), 318.
Cohen, T.R., Wolf, S.T., Panter, A.T., Insko, C.A., 2011. Introducing the GASP scale: a new measure of guilt and shame proneness. J. Person. Soc. Psychol. 100 (5), 947.
Cohn, A., Fehr, E., Maréchal, M.A., 2014. Business culture and dishonesty in the banking industry. Nature 516 (7529), 86–89.
Cojoc, D., Stoian, A., 2014. Dishonesty and charitable behavior. Exp. Econ. 17 (4), 717–732.
Conway, P., Peetz, J., 2012. When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or compensatory behavior. Person. Soc. Psychol. Bull. 38 (7), 907–919.
Dana, J., Weber, R.A., Kuang, J.X., 2007. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Econ. Theory 33 (1), 67–80.
Dolan, P., Galizzi, M.M., 2015. Like ripples on a pond: behavioral spillovers and their implications for research and policy. J. Econ. Psychol. 47, 1–16.
Effron, D.A., Bryan, C.J., Murnighan, J.K., 2015. Cheating at the end to avoid regret. J. Person. Soc. Psychol. 109 (3), 395.
Effron, D.A., Cameron, J.S., Monin, B., 2009. Endorsing Obama licenses favoring whites. J. Exp. Soc. Psychol. 45 (3), 590–593.
Effron, D.A., Miller, D.T., Monin, B., 2012. Inventing racist roads not taken: the licensing effect of immoral counterfactual behaviours. J. Person. Soc. Psychol. 103 (6), 916.
Effron, D.A., Monin, B., Miller, D.T., 2013. The unhealthy road not taken: licensing indulgence by exaggerating counterfactual sins. J. Exp. Soc. Psychol. 49 (3), 573–578.
Freedman, J.L., Fraser, S.C., 1966. Compliance without pressure: the foot-in-the-door technique. J. Person. Soc. Psychol. 4 (2), 195.
Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise–an experimental study on cheating. J. Eur. Econ. Assoc. 11 (3), 525–547.
Gino, F., Desai, S.D., 2012. Memory lane and morality: how childhood memories promote prosocial behaviour. J. Person. Soc. Psychol. 102 (4), 743.
Gino, F., Norton, M.I., Ariely, D., 2010. The counterfeit self: the deceptive costs of faking it. Psychol. Sci. 21 (5), 712–720.
Gino, F., Schweitzer, M.E., Mead, N.L., Ariely, D., 2011. Unable to resist temptation: how self-control depletion promotes unethical behaviour. Org. Behav. Hum. Decis. Process. 115 (2), 191–203.
Gneezy, U., 2005. Deception: the role of consequences. Am. Econ. Rev. 95 (1), 384–394.
Gneezy, A., Imas, A., Brown, A., Nelson, L.D., Norton, M.I., 2012. Paying to be nice: consistency and costly prosocial behaviour. Manag. Sci. 58 (1), 179–187.
Gneezy, U., Imas, A., Madarász, K., 2014. Conscience accounting: emotion dynamics and social behaviour. Manag. Sci. 60 (11), 2645–2658.
Gneezy, U., Kajackaite, A., Sobel, J., 2017. Lying aversion and the size of the lie. Am. Econ. Rev.
Gneezy, U., List, J., 2014. The Why Axis: Hidden Motives and the Undiscovered Economics of Everyday Life. Random House.
Gneezy, U., Rockenbach, B., Serra-Garcia, M., 2013. Measuring lying aversion. J. Econ. Behav. Org. 93, 293–300.
Gneezy, U., Rustichini, A., 2000a. A fine is a price. J. Legal Stud. 29 (1), 1–17.
Gneezy, U., Rustichini, A., 2000b. Pay enough or don't pay at all. Q. J. Econ. 115 (3), 791–810.
Greene, J.D., Paxton, J.M., 2009. Patterns of neural activity associated with honest and dishonest moral decisions. Proc. Natl. Acad. Sci. 106 (30), 12506–12511.
Hauser, O., Kraft-Todd, G., Rand, D., Nowak, M., Norton, M.I., 2016. Invisible inequality leads to punishing the poor and rewarding the rich. In: Academy of Management Proceedings, Vol. 1, p. 13841.
Hilbig, B.E., Thielmann, I., 2017. Does everyone have a price? On the role of payoff magnitude for ethical decision making. Cognition 163, 15–25.
Hollander, E.P., 1958. Conformity, status, and idiosyncrasy credit. Psychol. Rev. 65 (2), 117.
Horton, J.J., Chilton, L.B., 2010, June. The labor economics of paid crowdsourcing. In: Proceedings of the 11th ACM Conference on Electronic Commerce. ACM, pp. 209–218.
Horton, J.J., Rand, D.G., Zeckhauser, R.J., 2011. The online laboratory: conducting experiments in a real labor market. Exp. Econ. 14 (3), 399–425.
Jiang, T., 2013. Cheating in mind games: the subtlety of rules matters. J. Econ. Behav. Org. 93, 328–336.
Jordan, A.H., Monin, B., 2008. From sucker to saint: moralization in response to self-threat. Psychol. Sci. 19 (8), 683–689.
Jordan, J., Mullen, E., Murnighan, J.K., 2011. Striving for the moral self: the effects of recalling past moral actions on future moral behaviour. Person. Soc. Psychol. Bull. 37 (5), 701–713.
Kajackaite, A., Gneezy, U., 2017. Incentives and cheating. Gam. Econ. Behav. 102, 433–444.
Kajackaite, A., & Gneezy, U. (2015). Lying costs and incentives. UC San Diego Discussion Paper, UC.
Kessler, J.B., Milkman, K.L., 2016. Identity in charitable giving. Manag. Sci 64 (2), 845–859.
Kettle, S., Hernandez, M., Sanders, M., Hauser, O., Ruda, S., 2017. Failure to CAPTCHA attention: Null results from an honesty priming experiment in guatemala. Behav. Sci. 7 (2), 28.
Kouchaki, M., Gino, F., 2016. Memories of unethical actions become obfuscated over time. Proc. Natl. Acad. Sci. 113 (22), 6166–6171.
Landers, R.N., Behrend, T.S., 2015. An inconvenient truth: arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. Ind. Org. Psychol. 8 (2), 142–164.
Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honest people: a theory of self-concept maintenance. J. Market. Res. 45 (6), 633–644.
Merritt, A.C., Effron, D.A., Fein, S., Savitsky, K.K., Tuller, D.M., Monin, B., 2012. The strategic pursuit of moral credentials. J. Exp. Soc. Psychol. 48 (3), 774–777.
Merritt, A.C., Effron, D.A., Monin, B., 2010. Moral self-licensing: When being good frees us to be bad. Soc. Person. Psychol. Comp. 4 (5), 344–357.
Miller, D.T., Effron, D.A., 2010. Chapter three-psychological license: when it is needed and how it functions. Adv. Exp. Soc. Psychol. 43, 115–155.
Monin, B., Miller, D.T., 2001. Moral credentials and the expression of prejudice. J. Person. Soc. Psychol. 81 (1), 33.
Potters, J., Stoop, J., 2016. Do cheaters in the lab also cheat in the field? Eur. Econ. Rev. 87, 26–33.
Raihani, N.J., Mace, R., Lamba, S., 2013. The effect of $1, $5 and $10 stakes in an online dictator game. PloS One 8 (8), e73131.
Rudman, L.A., 2004. Social justice in our minds, homes, and society: the nature, causes, and consequences of implicit bias. Soc. Just. Res. 17 (2), 129–142.
Sezer, O., Gino, F., Bazerman, M.H., 2015. Ethical blind spots: explaining unintentional unethical behaviour. Curr. Opin. Psychol. 6, 77–81.
Shalvi, S., De Dreu, C.K., 2014. Oxytocin promotes group-serving dishonesty. Proc. Nat. Acad. Sci. 111 (15), 5503–5507.
Shalvi, S., Eldar, O., Bereby-Meyer, Y., 2012. Honesty requires time (and lack of justifications). Psychol. Sci. 23 (10), 1264–1270.

Shalvi, S., Handgraaf, M.J., De Dreu, C.K., 2011. Ethical manoeuvring: why people avoid both major and minor lies. Br. J. Manag. 22 (s1).

Shu, L.L., Gino, F., Bazerman, M.H., 2011. Dishonest deed, clear conscience: when cheating leads to moral disengagement and motivated forgetting. Person. Soc. Psychol. Bull. 37 (3), 330–349.

Starmer, C., Sugden, R., 1991. Does the random-lottery incentive system elicit true preferences? An experimental investigation. Am. Econ. Rev. 81 (4), 971–978.

Truelove, H.B., Carrico, A.R., Weber, E.U., Raimi, K.T., Vandenbergh, M.P., 2014. Positive and negative spillover of pro-environmental behavior: an integrative review and theoretical framework. Glob. Environ. Change 29, 127–138.

Von Hippel, W., Trivers, R., 2011. The evolution and psychology of self-deception. Behav. Brain Sci. 34 (01), 1–16.

Zhang, T., Fletcher, P.O., Gino, F., Bazerman, M.H, 2015. Reducing Bounded Ethicality: How to Help Individuals Notice and Avoid Unethical Behavior, Special Issue on Bad Behavior. Organizational Dyn. 44 (4), 310–317.

Zhong, C.B., Liljenquist, K., 2006. Washing away your sins: threatened morality and physical cleansing. Science 313 (5792), 1451–1452.

Zhong, C.B., Liljenquist, K.A, Cain, D.M., 2009. Moral self-regulation: Licensing and compensation. In: De Cremer, D. (Ed.), Psychological perspectives on ethical behavior and decision making. US: Information Age Publishing, pp. 75–89.

Chapter 6: Manuscript of Project 3

Rahwan, Z., Fasolo, B., Hauser, O.P., Deception: The Role of Experimenter Disclosure in Measuring Honesty (in preparation for submission for publication)

Zusammenfassung auf Deutsch

Die Verwendung von Täuschung in der Forschung ist entlang disziplinärer Linien gespalten, wobei typischerweise Psychologen dafür und Ökonomen dagegen sind. Ein Argument, das zugunsten der Täuschung vorgebracht wird, ist, dass es den Wissenschaftlern erlaubt, den wahren Zweck der Forschung gegenüber den Teilnehmern zu verschleiern, wodurch "Experimentator-Nachfrage-Effekte" reduziert und unverzerrte Messungen des Verhaltens erhalten werden. Die gegenteilige Ansicht besagt, dass Täuschung die Teilnehmer dazu bringen könnte, die "Spielregeln" in Frage zu stellen und das Verhalten in laufenden und nachfolgenden Studien zu beeinflussen. Hier testen wir, inwieweit die Offenlegung des Versuchszwecks - die häufigste Form der Täuschung - das Ehrlichkeitsverhalten der Teilnehmer beeinflusst. In zwei vorregistrierten Studien mit mehr als 2.000 Versuchsteilnehmern auf einer weit verbreiteten Online-Plattform finden wir heraus, dass die falsche Angabe des Versuchszwecks keinen Einfluss auf anreizgesteuerte Messungen der Ehrlichkeit hat. In Studie 1 variierte die Ehrlichkeit, die in einer Münzwurfaufgabe gemessen wurde, nicht, ob ein wahrer, unvollständiger oder falscher Zweck des Experiments angegeben wurde. In Studie 2 fügten wir eine Bedingung mit einem "absurden" falschen Zweck hinzu, um absichtlich Verdacht zu erregen, und verwendeten eine andere beliebte Ehrlichkeitsaufgabe, einen incentivierten Würfelwurf. In Übereinstimmung mit Studie 1 variierte die Ehrlichkeit nicht zwischen den Bedingungen und wurde nicht durch den Verdacht auf Täuschung beeinflusst. Unsere robusten Ergebnisse stimmen weitgehend mit einer früheren Studie (Gallo, Smith und Mumford 1973) überein, in der die Auswirkungen der Offenlegung des Experimentators auf die Konformität untersucht wurden. Wir untersuchen außerdem die Vorerfahrungen der Teilnehmer und ihre Einstellung zu Täuschungen sowie die wahrgenommene Toleranz unter anderen Teilnehmern. Die selbstberichtete Erfahrung mit Täuschung vor diesem Experiment sagte das Verhalten in unseren Ehrlichkeitsaufgaben nicht voraus. Während 78% der Teilnehmer persönlich nichts gegen die Verwendung von Täuschungen einzuwenden haben, glauben sie fälschlicherweise, dass nur 54% ihrer Mitspieler Täuschungen gegenüber tolerant sind. Außerdem erwartete etwa ein Viertel der Teilnehmer, dass sich ihr Verhalten in zukünftigen ähnlichen Aufgaben ändern würde, nachdem sie der Täuschung ausgesetzt waren. Insgesamt deuten unsere experimentellen Ergebnisse darauf hin, dass die Exposition gegenüber einem falsch angegebenen Versuchszweck die Messung der Ehrlichkeit bei Würfelwürfen und Münzwürfen nicht verzerrt, obwohl die Verwendung dieses Zwecks bei den Teilnehmern nicht ohne Vorbehalte ist.

Deception: The Role of Experimenter Disclosure in Measuring Honesty

Zoe Rahwan, Barbara Fasolo, Oliver P. Hauser

Abstract

The use of deception in research is divisive along disciplinary lines, with typically psychologists in favor and economists opposed. One argument put forward in favor of deception is that it allows scholars to disguise the true purpose of the research to participants, reducing "experimenter demand effects" and obtaining unbiased measures of behavior. The opposing view holds that deception could lead participants to question the "rules of the game" and affect behavior in current and subsequent studies. Here, we test to what extent disclosing experimental purpose—the most common form of deception—affects participants' honesty behavior. Across two pre-registered studies with more than 2,000 experimental participants on a widely-used online platform, we find that falsely stating the experimental purpose has no impact on incentivized measures of honesty. In Study 1, honesty measured in a coin flipping task did not vary whether a true, incomplete or false purpose of the experiment was given. In Study 2, we added a condition with an 'absurd' false purpose to deliberately evoke suspicion and used another popular honesty task, an incentivized die roll. Consistent with Study 1, honesty did not vary across conditions and was unaffected by suspicion of deception. Our robust findings are broadly consistent with an earlier study (Gallo, Smith, and Mumford 1973), which studied the effects of experimenter disclosure on conformity. We further study participants' prior experience and attitudes about deception, as well as perceived tolerance among other participants. Self-reported exposure to deception prior to this experiment did not predict behavior in our honesty tasks. While 78% of participants do not personally object to the use of false purpose, they falsely believe that only 54% of their peers are tolerant of deception. Furthermore, around a quarter of participants expected that participants' behavior would change in future similar tasks after being exposed to deception. Overall, our experimental findings suggest that exposure to a falsely stated experimental purpose does not bias the measurement of honesty using die rolls and coin flips, though its usage is not without some reservation among participants.

Keywords experimental methods, meta-science, deception, false purpose, experimenter disclosure, honesty, open data, open materials, pre-registered

> "…there is a world of difference between not telling subjects things and telling them the wrong things. The latter is deception, the former is not." — John D. Hey (1998)

> "… in my long experience, the majority of subjects care more about their time and their money or other incentives than they care about any duplicity being undertaken against them. Student populations, in particular, find these practices no more loathsome than lack of transparency in grading…" — McDermott (2013)

The use of deception has long been controversial in social sciences and prevalent in prestigious psychology journals in the post World War II period (Hertwig and Ortmann 2008a). Revelations of abuse of human subjects in deceptive experiments in both wartime (e.g. Nazi medical trials (Roelcke 2004), US human radiation exposure trials (Advisory Committee on Human Radiation Experiments 1996) and peacetime (Tuskegee Syphilis Experiment (Schuman et al. 1955), Milgram's obedience experiments (1963), Zimbardo's prison experiment (1971)) prompted major reforms in ethical oversight of scientific experiments involving human subjects. In the US, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was established in 1974. This commission subsequently published The Belmont Report ("The Belmont Report" 1979) which contains a guiding set of principles for Internal Review Boards

(IRBs) or ethics committees, which endure until today both in the US and beyond (Raymond 2019).

Divisive views between research communities

Economists generally prohibit the practice of deception. Objections to deception seem to be motivated by dominantly utilitarian rather deontological reasons (Roth 2001; Barrera and Simpson 2012), notably that of invalidating tests of economic theory. Ariely and Norton (Ariely and Norton 2007) speculate that the categorical avoidance of deception is motivated by traditional economic axioms of behavior. Specifically, full and honest information on the rules of the game is required to enable participants to make an accurate utility-maximising decision. Deception may cause participants to doubt the veracity of information and materials presented by the experimenter (Cooper 2014; Jamison, Karlan, and Schechter 2008), undermining the ability to accurately test hypotheses. Despite these concerns, there is negligible recent experimental evidence to assess whether deception affects behavior, with the notable exception of Jamison et al (2008) and to a lesser extent, Krasnow et al. (2020) who explored differences in behavior and suspicion across subject pools from MTurk, economics and psychology laboratories.

Another concern expressed by economists regards future experimental participation. The break in trust between the experimenter and participants from deception has been argued to reduce the willingness for future participation in experiments. Some evidence has emerged for a gendered selection bias into future experiments after having been exposed to deception (Jamison, Karlan, and Schechter 2008).

In contrast, psychologists argue that deception offers a means to study important yet uncomfortable aspects of the human condition (e.g. conformity, obedience) (Bortolotti, Mameli, and Mameli 2006). Without a means to deceive participants, particularly regarding the nature of the study, it is argued that measures would be biased. Some have gone so far as to argue that the use of experimenter deception can support efforts to create a "more efficient and just society" (Bortolotti, Mameli, and Mameli 2006). However, the frequency of use of deception (Hertwig and Ortmann 2008a; Adair, Dushenko, and

Lindsay 1985) suggests that deception is not necessarily only reserved for studies of particular importance for society.

In fact, some psychologists have suggested that deception does not lead to an aversive experience for participants. Indeed, Christensen (1988) concludes that participants in deceptive experiments “enjoyed the experience more, received more educational benefit from it, and did not mind being deceived or having their privacy invaded.”

Institutional and community norms

Independent of the source of the concern which leads economists to avoid the use of deception and psychologists to accept it, there are significant, tangible consequences for researchers engaged in the practice. Perhaps most notable is the immediate rejection of research articles using deception from economic journals. In addition, more interdisciplinary journals, such as the *Proceedings of the National Academy of Sciences*, as well as grant funding bodies may reject research that contains or proposes to use deception (McDermott 2013; Jamison, Karlan, and Schechter 2008; Cooper 2014; K. S. Cook and Yamagishi 2008). Together, these barriers may pose tangible barriers to career advancement (Barrera and Simpson 2012).

There also exists institutionalized discouragement in the use of deception, even among psychologists. Professional bodies, such as the American Psychologists Association (APA), and many, though not all university IRBs, discourage— do not ban—deception. Commonly, university IRBs require approval and justification to use deception in social science experiments (K. S. Cook and Yamagishi 2008). Notably, APA Guidelines (“Ethical Principles of Psychologists and Code of Conduct” 2016) require a debriefing and permission for participants to withdraw from the study after learning of the deception. However, Adair et al (1985) find evidence that the latter is not always being implemented. Despite these arrangements, the common usage of deception among psychologists suggests that they have provided limited deterrence.

Forms of deception

Deception comes in many forms. In keeping with the literature (Hertwig and Ortmann 2008b), we define deception as the intentional and explicit misleading of participants, as opposed to withholding information on hypothesis or experimental manipulations. Further, we draw upon—and in some parts update to reflect contemporary research tools and platforms (*see Appendix 1*)—Sieber et al.'s (1995) taxonomy, in which they identify eight types of deception, including false purpose, giving false feedback to participants and not disclosing to participants that they are part of a study.

We focus on the use of deception regarding the true purpose of study. 'False purpose' is defined as participants "may be given or be caused to hold, false information about the main purpose of the study" (Sieber, Iannuzzo, and Rodriguez 1995). An example would be claiming that a study is about "Life and Satisfaction" (Cohn, Fehr, and Marechal 2014) when it is really about how professional culture affects honesty. These forms of deceptions are widespread. In fact, false purpose deception appears to be the most common form of deception in psychology. A survey of 177 studies from 59 pieces published in the Journal of Experimental Social Psychology published in 2002, found that 87% disclosed a false purpose to participants (Hertwig and Ortmann 2008a). Given this prevalence, understanding and exploring empirically what—if any—effect this form of deception has on participant behavior is important.

Variations in the disclosed experimental purpose could affect behavior in different directions: participants could succumb to experimenter demand or retaliate (Orne 1962). Past research on this question is limited. While some research (e.g. (Boynton, M. H., Portnoy, D. B., & Johnson, B. T. 2013; Gerdes 1979) combined false purpose with other types of deception, Gallo et al. (1973) is the only study we found which experimentally manipulates deception solely in the form of false purpose.

Gallo et al.'s (1973) experiment (n=120) was conducted among (female) first-year psychology students undertaking a conformity task. Participants were either informed that it was a "distance perception" study (false purpose), or a distance perception study in which researchers also had an interest in conformity (partial information) or conformity (true purpose). No differences were found in the means or variances of the conformity tasks across the conditions. Notably, over half of participants in the 'false purpose' condition suspected that the experiment was actually about conformity, and that "quite a few" participants in the 'true purpose' also questioned the actual purpose (Gallo et al. (1973)). The authors drew two conclusions from their studies regarding the use of deception; (i) even those participants with little-to-no experimental experience, were not that 'naive' and (ii) little attention may be given to the stated experimental purpose.

More recently, Krasnow (2020) explored differences in behavior in common economics games between different subject pools in North America. Their samples encompass two on-campus economics laboratories (which disallowed deception), two on-campus psychology laboratories (which permitted deception), and the online labor market Amazon Mechanical Turk (where deception is possible, see (Mason and Suri 2012). While they found differences in behavior across the three different subject pools, this could not be solely attributed to deception.[1] Rather, the differences may have simply arisen from selection biases found in each of the populations and/or the tasks used. Measures of 'spontaneous' (i.e. unprompted) suspicion of deception did not differ between economics and psychology labs, while within the MTurk sample, suspicion did not vary with self-reported past exposure to deception. For 'prompted' suspicion of deception measures, psychology subject pools were more suspicious of being deceived and more confident in their suspicion than participants from economics laboratories or MTurk.

Another relevant consideration is how participants may view false purpose, in terms of experimenter obligations and its effect on them. An earlier survey (Epstein, Suedfeld, and Silverstein 1973) found that a minority (20-24%) of participants felt that experimenters were obligated to inform the participant of the (presumably, truthfully) purpose of the experiment. A failure to do so was categorised as "slightly undesirable" though not

[1] *Participants' individual historical exposure to deception was not coded for in the psychology labs. That is, it is not certain that participants had been deceived, how frequently they had been deceived, what type of deception they had been exposed to, the last time they were deceived, etc. Similarly, for MTurk workers, their prior exposure to deception was not elicited in the survey, nor is recorded by the platform.*

egregiously so, unlike causing embarrassment or providing unclear instructions (Epstein, Suedfeld, and Silverstein 1973).

Current research

Here, we use experimental tools to better understand how the type of disclosure regarding experimental purpose (true, incomplete disclosure or false) affects behavior in two common honesty tasks on a popular experimental platform, Amazon Mechanical Turk (MTurk) (Bohannon 2016). We also elicit incentivized views on norms regarding experimental use of false purpose, and non-incentivized views on deception in general and its perceived spillovers on participants.

In addressing the studying whether experimenter disclosure affects honesty, we pursued three research questions. Firstly, we were interested to learn whether disclosing the true purpose of the experiment (honesty) would increase honest behavior relative to other conditions. One might expect an increase due to "experimenter demand effects" whereby participants act in a way that they believe is socially appropriate and expected by the experimenter (Zizzo 2010). Indeed, in a large-scale meta-analysis, Abeler et al (2019) found evidence that a sense of being observed by an experimenter reduces dishonesty.

Secondly, does a deceptive stated experimental purpose increase or reduce dishonesty, relative to other stated purposes? In a meta-analysis of 565 experiments, Gerlach et al (2019) find the use of deception to be associated with less dishonesty, dominantly in sender-receiver games (k=165), wherein at a minimum, participants were mis-led by experimenters regarding the existence of another participant. However, the presence of deception was not associated with changes in the level of honesty in the other three tasks assessed: coin-flipping tasks (k=163), die roll tasks (k=129), and matrix tasks (k=101).

Thirdly, does the nature of deception evoke different levels of honesty? That is, if a neutrally framed false purpose does not lead to changes in behavior, would a provocative—even absurd—false purpose arouse more suspicion and alter behavior? Gerlach et al (2019) speculate that association between deceptive experimental settings and lesser dishonesty could be caused by suspicion which, in turn, motivates a reduction in socially undesirable behavior. This is consistent with (Silverman, Shulman, and Wiesenthal 1970) who found, in a relatively low-powered study with 98 participants across three conditions, that participants in deception conditions were more likely to present themselves favorably and (Stricker, Messick, and Jackson 1967) which found a correlation between suspicion and social desirability response style. Suspicion may also prompt some form resistant response strategy, as opposed to succumbing to experimenter demand (Orne 1962). Both Stricker et al. (1967) and Glinski (1970) reported an association between suspicion of deception and reduced conformity. Meanwhile, more recent research found no link between suspicion and behavior on common experimental economics tasks (Krasnow, Howard, and Eisenbruch 2020). It is unclear whether in honesty experiments, suspicion would lead to more (due to a resistant response strategy) or less (due to social desirability concerns) or no changes in honesty.

Statement of Relevance

False purpose is the most commonly used form of deception in psychology experiments. Some argue that its usage is critical to yielding results, unbiased by social desirability, when measuring important, yet uncomfortable human behaviors. Others, commonly economists, argue that it may 'taint' participant pools, causing biased results in subsequent studies, due to experimenters breaching participant trust and selective participant attrition. Another concern is that the use of deception could provoke suspicion and consequently bias behavior in the study deploying deception. Despite strong views, no work has been published on the effects of false purpose deception since the 1970s. Here, we explore for the first time, whether the type of disclosed experimental purpose (true, incomplete disclosure or false) affects the measurement of a commonly measured behavior, honesty. We found among large samples on a popular online experimental platform across two commonly used honesty tasks that participant behavior was insensitive to the form of stated experimental purpose. This held even where suspicion was provoked by using an absurdly deceptive experimental purpose (a study that claimed to be about "Juggling Clowns"). While participants were generally tolerant of researcher use of false purpose, an

> incentivized norm measure revealed that much less tolerance was expected among peers. While these findings suggest that the use of false purpose deception may not affect these experimental measures of honesty, more research is needed to understand the effects on other measures and tasks, any adverse spillovers on future experimental behavior and the effects of other types of deception.

We drew upon two large samples from a popular platform for running experiments, Amazon's Mechanical Turk (MTurk), using two commonly used honesty tasks (Gerlach, Teodorescu, and Hertwig 2019) - die rolls and coin flips. Pilots were run for each study (*n*=30, and *n*=51 respectively) for the purposes of discovering any errors in the survey and for determining the average time taken to complete the survey. Subsequently, we ran two highly-powered (pooled n=2,134) and pre-registered Studies 1 and 2.

Study 1 experimentally varied three types of disclosed experimental purpose: true (participants were told the study is about "Honesty"), incomplete disclosure ("Norms and Attitudes") and deception ("Life and Satisfaction"). In addition to measuring the incentivized, self-reported outcome from a 10-round coin flipping task (our main, pre-registered outcome variable of interest), we also tested for differences across conditions in the general suspicion of being mis-led, and if a suspicion was held, we asked about the nature of the suspicion (free text response).

In Study 2, we used an incentivized, self-reported die roll task and similarly varied the types of disclosure regarding experimental purpose across four conditions: true ("Honesty"), incomplete disclosure ("Judgement and Decision Making"), standard deception ("Life and Satisfaction") and, as an additional condition not present in the previous study, 'absurd deception' (where participants were told the experiment was about "Juggling Clowns"). This condition was added to provoke suspicion and observe its effect on honesty. We amended the text for the incomplete disclosure condition from Study 1 to improve its ecological validity by using another commonly used wording. In other key changes from Study 1, we undertook manipulation checks and incentivized measures of honesty norms and norms regarding the use of false purpose. We refined the suspicion measure to include a measure of confidence in the belief of being deceived. We also collected participant views regarding different types of deception and expected spillovers on behavior and attitudes from exposure to deception.

Study 1

Method

Our data, materials, and preregistration are available on the Open Science Framework (OSF): https://osf.io/f6gmb/?view_only=2ad7305cce094ff4a349850dcbcc304e . Approval for this research (including the use of deception) was provided by the London School of Economics Research Ethics Committee (reference #000582). Following the conclusion of the study, all participants were fully de-briefed via a message sent separately on the MTurk messaging service.

Participants

This study was run on 27 July 2017 on Amazon Mechanical Turk using US participants. We strived to recruit 900 participants across three conditions (i.e. 300 participants per condition). The sample size was determined from previous studies conducted using the same 10-round coin flipping task, to enable the detection of a 0.5 difference in total reported coin flip wins between two conditions, the minimum difference to result in a change in payoff. In the end, a total of 927 participants completed the survey (mean age = 36 years, age range = 18-76, 423 men, 498 women, 4 identifying as ‘other’).

Materials and procedure

This initial study was motivated by differences in stated experimental purpose in undertaking a replication (Z. Rahwan, Yoeli, and Fasolo 2019) of a highly-cited field study (Cohn, Fehr, and Marechal 2014; “Altmetric – Business Culture and Dishonesty in the Banking Industry” 2020). The original study used a false purpose (“Life and Satisfaction)” while the replication studies used incomplete disclosure (“Norms and

Attitudes") (as a result of differing Research Ethics Committee standards and field partner constraints). Because the replication studies did not find the same results as the original study, one hypothesis was whether this may have been driven by differences in the disclosed experimental purpose between the studies. However, as the initial and replication studies were conducted at different times and different populations, it was not possible to isolate the role of the disclosed experimental purpose.

The present study therefore studies this question systematically by varying the stated purpose across three randomized conditions. When participants joined the study, they were presented with a welcome page which randomly presented one of three types of experimental purposes: true ("Honesty"), incomplete disclosure ("Norms and Attitudes") and false ("Life and Satisfaction"). After consenting to join the study, all participants were asked questions about happiness, satisfaction and leisure activities. We included these questions for all conditions so that even participants in the 'false' condition ("Life and Satisfaction") could plausibly believe that we, the researchers, were interested in these outcomes; our main interest was, of course, their behavior in the honesty task described below. (This design feature is also consistent with the original study we were attempting to replicate.)

Next, the key outcome variable—honesty—was measured in a 10-round "coin flipping" task. The multi-round coin flipping task works as follows (Cohn, Fehr, and Marechal 2014): In each round, participants could win a US5 cent bonus for reporting a winning coin flip. Participants were then asked to submit their own answer when asked what the coin flip revealed. This self-report meant that there was an opportunity to cheat: Participants were informed of the winning outcome ahead of submitting their answers, giving them an opportunity to cheat to receive up to a maximum payoff of US 50 cents if they cheated in every one of the ten rounds. Given the binomial distribution of a fair coin toss, we can detect both the presence of cheating and the difference in cheating across conditions.

We then took an implicit measure of moral feelings after the coin-flip task by asking participants to solve word fragment puzzles. Four of the six puzzles could be solved with ‘moral’ words (pure, virtue, moral and ethical), inspired by (Gino et al. 2011). We were interested to see if being able to spot the moral words varied across the experimental conditions. Participants also completed additional measures (as per Cohn et al., 2014) to capture self-reported materialism, altruism, competitiveness, other-regarding concerns in general and with regard to their work.

Basic demographic information was collected (age, gender, education, region of US residence, MTurk experience). At the end of the survey, we applied a ‘prompted’ measure of suspicion (Krasnow, Howard, and Eisenbruch 2020) to determine whether individuals were suspicious about being deceived in this study and about what they thought they were deceived about. Beliefs have been long-held (e.g. (Kelman 1967) that the use of deception can increase the suspicion of participants towards researchers. That said, Krasnow et al. (2020) did not find that suspicion was linked to past exposure to deception. We took measures of past perceived deception on MTurk and its nature, experience in coin-flipping tasks, recency of the last coin-flipping task and participants’ self-reported emotions after and the speculated purpose of completing such tasks.

Analysis

We conducted analysis using non-parametric measures, due to the skewed nature of our main measure of interest, i.e. the self-reported number of winning coin tosses. We also undertake linear regression analyses, modelling both total winning coin tosses and the reporting of a winning coin toss, clustering standard errors at the level of participants across rounds and controlling for a range of demographic and other variables. This is as per the pre-registration.

Results

Honesty

Firstly, we assess the presence of dishonesty across all conditions. In all conditions, we detect the presence of dishonesty relative to that which would be predicted by the theoretical distribution of a fair coin (using Wilcoxon one-sided tests: $W_{\text{TruePurpose}}$

=168,648, $W_{\text{IncompleteDisclosure}}$=173,090, $W_{\text{FalsePurpose}}$=177,626, all *p*s<0.001). Further, in all conditions the proportion of individuals reporting the maximal number, 10, of winning outcomes (9%, 9%, 10% in the true, incomplete and false purpose conditions, respectively), is well above that predicted by a fair coin (~0%). In stark contrast, at the other end of the extreme, no participants reported the minimal number of winning outcomes - zero (see Fig. 1a). For the entire sample, the average payoff is US29 cents (95% CI= [28, 30]), marking a forgoing of 42% of the maximal reward, US50 cents.

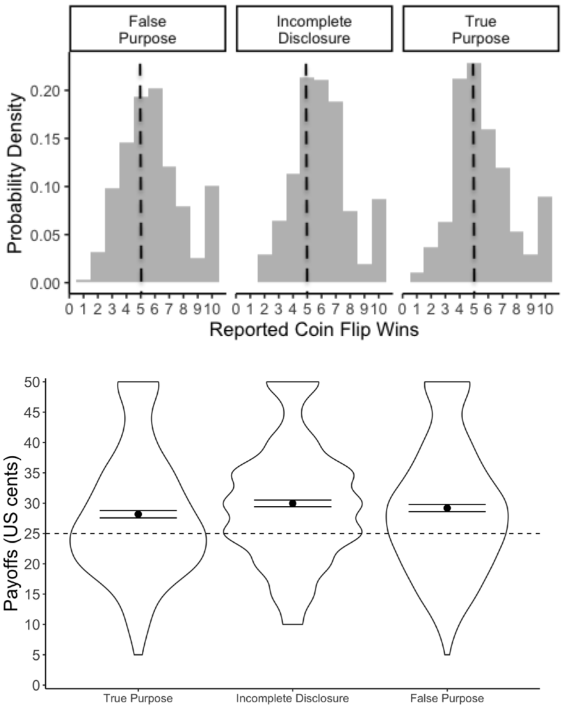


Fig. 1a. Distributions of reporting coin flip wins in Study 1: The probability densities of the range of reported outcomes from the coin flipping task (0-10) in each condition, which varies the nature of the stated experimental purpose (true, incomplete disclosure and false). In each condition, cheating is detected relative to a theoretical fair coin distribution, the mean of which is marked with a dotted line. **b.** Reflecting the cheating in all conditions, the average payoff is 25 US cents predicted by a theoretical distribution. While visually small differences in average cheating and average payoffs can be observed between conditions, these differences are economically minimal and rarely statistically significant. Error bars indicate standard errors of the mean.

Secondly, we assess whether there are differences in honesty across conditions. Using non-parametric tests, we find that there are differences across the conditions (*H(Kruskal-Wallis)*(2)=7.213, *p* =0.027) However, the only pairwise comparison of conditions in which a significant difference is detected is that of true purpose

($M_{\text{ReportedWins}}$=5.64) and incomplete disclosure ($M_{\text{ReportedWins}}$=5.99, $W_{two\text{-}sided}$=40,618, p=0.006, $p_{\text{critical-adjusted}}$=0.05/3 = 0.017(given multiple comparisons between the three conditions). The false purpose condition ($M_{\text{ReportedWins}}$=5.84) was not statistically distinct from the other conditions. Using a probit regression model with control variables and clustering at the level of individual participants over the multi-round task, no differences are detected across the conditions (*able S1*).

Third, we explored whether any overall increase or decrease in cheating across conditions could be mapped to an increase in transparency about the true purpose of the research (Fig. 1b). That is, as transparency increases from least transparent (i.e. false purpose condition) to medium transparent (i.e. incomplete disclosure) to fully transparent (i.e. true purpose), there are no significant effects on honesty (*JT*=148340, *p*=0.194).

Suspicion of Deception

Less than one-fifth of participants expressed suspicion of being deceived about any part of this study ($M_{\text{TruePurpose}}$=0.18, $M_{\text{IncompleteDisclosure}}$=0.16, $M_{\text{FalsePurpose}}$=0.18). This could under-represent actual suspicion (Taylor and Shepperd 1996). No differences were found between the conditions (*H*(2)=0.773, *p* =0.680). Notably, those who had previously been exposed to deception in the past, were ~17% more likely to expect to be deceived in the current experiment (*Table S4*). With reference to deception regarding the stated purpose of the current study, even less suspicion was identified. Among those expressing general suspicion, we manually coded what they thought they had been deliberately misled about in this and past studies. Only 9, 12 and 15 participants in the true purpose (3.0%, n=302), incomplete disclosure (3.9%, n=309) and false purpose (4.8%, n=316) conditions, respectively, believed they were deceived with regard to the stated purpose of this study.

Survey Control Variables

We found that most controls relevant to the experiment had no effect on the likelihood of reporting a winning outcome (see Table S1). With reference to the probit model, expectations of having been deceived in this or past experiments, experience with the coin-flipping tasks and the recency of any past coin flipping task experience had no effects. However, a measure of participant experience on the MTurk platform in number of years (mean = 1.99 median = 1.0, range=(0,10)) did have an impact on reporting more winning coin tosses, though the effect was small: for each additional year of MTurk experience, a participant was 1% more likely to report a win.

Demographic Control Variables

We find that demographics do have some influences, albeit small, on the probability of reporting winning coin tosses (*see Table S1*). Again, with reference to the probit model, we find that for each additional year of age, the probability of reporting a winning coin toss declines 0.2% (p<0.001) and that men are 3% more likely to report a winning coin toss relative to women and other gender identities (p=0.015). Differences are found across US regions; relative to the Mid-West, participants in the South were less likely to report a winning coin toss (p<0.024). Higher education was found to have no effect on reported winning coin tosses.

Study 2

Our data, materials, and preregistration are available on the OSF: https://osf.io/f6gmb/?view_only=2ad7305cce094ff4a349850dcbcc304e. Approval for the use of deception was provided by the London School of Economics Research Ethics Committee (reference #000921). Following the conclusion of the study, all participants were fully de-briefed within the survey.

Method

Participants

Study 2 was launched on March 9, 2021 on Amazon Mechanical Turk using US participants. We strived to recruit 1,200 participants (i.e. 300 per condition) in line with the first experiment. This also ensured statistical power to experiments using the same die roll task (Fischbacher and Föllmi-Heusi 2013; Kajackaite and Gneezy 2017). In the end, 1,209 participants completed the survey (mean age = 40 years, age range = 18-79, 556 men, 636 women, 10 non-binary individuals, and 7 preferring not to answer).

Materials and procedure

We extended Study 1 to explore whether (i) deliberately provoking suspicion via stated experimental purpose (the fourth condition claimed to be about "Juggling Clowns") can affect honesty behavior, and (ii) whether inattention or lack of seriousness in completing the survey could explain the insensitivity to stated experimental purpose.

Similar to Study 1, we recruited participants from MTurk, posting that the study was about 'judgement and decision making.' When participants joined the study, they were presented with a welcome page which randomly presented one of four types of experimental purposes: true, incomplete disclosure and two false conditions (either a standard false purpose: "Life and Satisfaction"; or an absurd purpose: "Juggling Clowns"). On the second page, participants were presented with a standard consent form which reiterated the stated experimental purpose.

In keeping with Study 1, participants first completed questions about life and satisfaction, followed by the honesty task. In Study 2, however, we chose a slightly different measure of honesty to ensure that our earlier results were not simply an artifact of the honesty measure used: in Study 2, we therefore used a one-shot die-roll with variable payoffs (Fischbacher and Föllmi-Heusi 2013). Participants were asked to use a real (physical) die or visit a die-rolling site, and roll the die to ensure that it is fair (Shalvi et al. 2011). Once satisfied, participants were asked to roll the die and self-report the die roll outcome. Participants were aware of the variable bonus depending on their self-report: US10 cents for reporting 1, US20 cents for 2, ...US50 cents for 5, and US0 cents for 6. Given the equal probability of each outcome, at a group level, comparison can be made to a uniform distribution to assess the presence and degree of dishonesty, but dishonesty cannot be identified at an individual level.

Next, we assessed the presence of suspicion again using ‘prompted’ measures from Krasnow (2020). Ahead of this, we provided reassurances that answers would not affect compensation, as per Blackhart el al. (2012). Participants were asked if they believed “they were intentionally misled about any part of this study” (yes/no), and how confident they were in their belief (7-point scale, with 1 = “I am positive I was not deceived” and 7 = “I am positive I was deceived”). We then asked what they believed that had been deceived about (free text response).

Participants were then asked to recall the stated purpose of the study as our manipulation check. They were offered a US10 cent bonus for selecting the correct response from a list which noted the four conditions and “I don’t know.” For participants who did not choose “I don’t know,” we asked them what they thought the true purpose of the study was, offering the following options: the same as the stated purpose, “I don’t know”, or “Other” (free text response). Of the 282 participants choosing ‘Other’, 247 (88%) correctly reported a belief that we were assessing honesty.

Next, we assessed individual judgements and peer expectations with regards to researcher use of false purpose on MTurk. We asked whether individuals thought it should be permissible for researchers to use deception in studies (yes/no) and how many (out of 100 MTurker peers) they thought would believe it to be permissible (participants chose from 11 increments from 0-100). A reward of US10 cents was offered for the correct answer (rounded to the nearest 10) to further mitigate social desirability effects (Bicchieri 2016).

Participants were then asked to complete four questions regarding trust in researchers and science in general. Specifically, we asked how much participants trusted researchers they had previously worked for on MTurk, science in general, that we and other researchers would pay promised bonuses.

Next, we introduced a section regarding experience, concerns and expected spillovers from deception in general. We adapted the definitions of deception from Sieber et al (1995) - collapsing the 'unaware of measure' and 'unaware of participation' measures, and updated other measures to reflect recent research practices (see Appendix 1). We then asked if participants recalled being debriefed regarding a deception (yes/no/unsure). For those answering 'yes' or 'unsure', we probed which type of deception they had been exposed to.

All participants were asked about their level of concern - philosophical or practical - about the seven different types of deception (5-point scale; "not at all concerned", "moderately concerned", "very concerned"). Based on previous exposure to deception, we probed what spillovers they had either experienced or would anticipate to experience; behavior in similar tasks in future studies (change/no change), behavior in different tasks in future studies (change/no change), level of trust in researchers and science, willingness to participate in future studies, level of attention, seriousness and suspicion in future studies (decreased, no change, increased). Open text questions for other effects from deception and thoughts about researchers' use of deception were then posed.

We asked about die roll experience (0-100, more than 100 tasks), what participants thought the purpose of the die-rolling task was (open text). After providing a reassurance that it would not affect any payments (Blackhart et al. 2012), we asked how serious participants were in undertaking this survey (5-point scale, anchored with "not at all" and "very serious.")

Finally, we collected demographic information (age, gender, education, relative income, income, political and religious preferences (Huang et al. 2021)), and data related to MTurk experience (number of year, number of HITs, share of academic study HITs, importance of their work on MTurk for income and for generating a sense of purpose (5-point scale, anchored with "not at all important" and "very important").

Analysis

We conducted analyses using tools outlined in Study 1. That is, we used non-parametric tests, due to the skewed nature of our main measure of interest, i.e. the die roll outcome. The die roll outcome is presented in terms of payoffs for ease of comprehension. We also undertake linear regression analyses, modelling the payoff against treatment, controlling for a range of demographic and other variables. This is as per the pre-registration.

Results

Honesty

Similar to Study 1, we find the presence of dishonesty across all conditions (see Fig. 2). By comparing die roll outcomes to that of a fair die (i.e. a uniform distribution), dishonesty is detected ($d(\text{Kolmogorov-Smirnov, one-sided})_{\text{TruePurpose}}$ =0.32, $d_{\text{IncompleteDisclosure}}$=0.26, $d_{\text{StandardFalsePurpose}}$=0.26, $d_{\text{AbsurdFalsePurpose}}$=0.23, all *p*s<0.001).

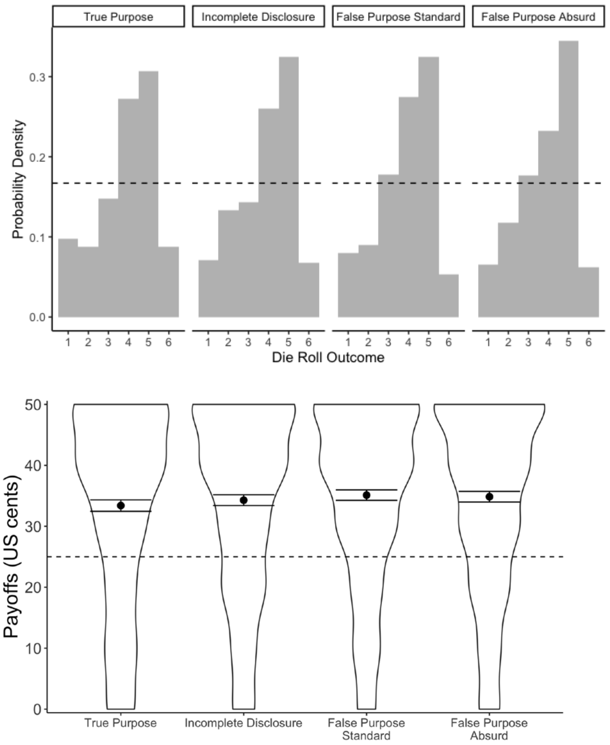


Fig. 2. a. Distributions of reported die rolls in Study 2. The probability density of the report outcome from the die rolling task (1-6), for each of the four conditions (true, incomplete disclosure and false - standard and absurd). The dashed line represents what would be expected in the absence of dishonesty -

a uniform distribution with each die roll outcome having a probability of ⅙. **b**. The ‘violins’ represent the distributions of the payoffs across each of the four conditions. Dots represent the average payoff for each condition. Error bars indicate standard errors of the mean. The dashed line represents the average outcome predicted by the theoretical distribution of a fair six-sided die with the variable pay-off structure we deployed.

Secondly, we assess whether there are differences in honesty across conditions. Using non-parametric tests, we find that there are no differences across the conditions ($H(3)=0.24$, $p=0.97$). This is also supported by linear regression models (*see Table S2*).

Thirdly, we conduct comparisons between conditions. We do not find evidence that stating the true purpose (i.e. measuring honesty) invokes more or less dishonesty relative to all other forms of experimental purpose ($W_{two\text{-}sided}$=130,324, p=0.314). We also find no influence of deceptive disclosures (both standard and absurd) in their effect from non-deceptive disclosures (both true purpose and incomplete disclosure - $W_{two\text{-}sided}$=176,816 p=0.317). Finally, the type of deceptive disclosure - standard or absurd - does not appear to generate differing levels of honesty when compared to each other ($W_{two\text{-}sided}$=45772, p=0.932).

Honesty Norms

On average, MTurk participants cheated, though far from maximally. The average payoff from the die roll task was US 34 cents against a maximum of US 50 cents, equating to 31% of the total reward being foregone, on average. When asked what participants expected of their MTurk peers, there was a clear expectation across all conditions of elevated cheating (see Fig. 3). Using the incentivized measure of social norms, we find that MTurk participants expected their colleagues to report outcomes that would lead to payoffs of US 44 cents, on average, equating to a lesser 13% of the total reward being foregone. Similarly, while 33% of participants engaged in maximal cheating in the die roll task themselves, on average they expected 68% of other MTurkers to do so. Thus, expectations of MTurker peer honesty was far worse than justified by the data in the task in this experiment. Peer expectations did not vary across conditions ($H(3) = 3.28$, p=0.35).

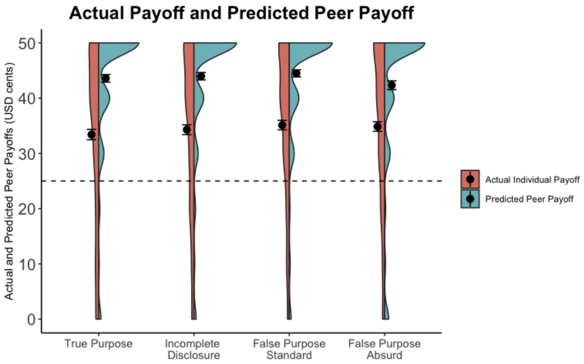


Fig. 3. Results from Study 2. The dashed line represents the average outcome predicted by the theoretical distribution of a fair six-sided die with the variable pay-off structure we deployed. Dots represent the average payoff for each condition for both what the individual reported in the die roll task (their actual behavior, see "Self" in green) and their incentivized expectations of what their peers would report (see "Other" in red) in the die roll task. Error bars indicate standard errors of the mean. The 'violin' presents the distribution of payoffs for each condition. The wider the violin, the greater the probability of the relevant (predicted) payoff.

Manipulation check

While the majority attended to the stated experimental purpose across the sample ($M_{TotalSample}=0.70$, 95% CI= [68%, 73%]), there was notable variation in the proportion of participants correctly identifying the stated experimental purpose ($M_{\text{TruePurpose}}$=0.67, 95% CI =[62%, 73%], $M_{\text{IncompleteDisclosure}}$=0.41, 95% CI = [54%, 65%], $M_{\text{StandardFalsePurpose}}$=0.71, 95% CI=[66%, 76%], $M_{\text{AbsurdFalsePurpose}}$=0.84, 95% CI =[79%, 88%]). Those in the 'absurd' false purpose condition were ~17% (p<0.001, 95% CI = [11%, 23%]) more likely to correctly identify the stated purpose of the experiment relative to the true purpose, while those in the incomplete disclosure condition were ~8% less likely to (p=0.037*,* 95% CI = [-15%, 0%] - see *Table S5*). No differences were identified between the true purpose and the standard false purpose conditions. The ability to correctly identify the stated experimental purpose was not associated with any differences in coin flip behavior or payoffs (see *Table S2*).

Suspicion of Deception

The proportion of participants holding suspicions of researcher deception varied across conditions ($M_{\text{TruePurpose}}$=0.19, 95% CI = [0.14, 0.23], $M_{\text{IncompleteDisclosure}}$=0.12, 95% CI = [0.08, 0.16], $M_{\text{StandardFalsePurpose}}$=0.14, 95% = [0.10, 0.18], $M_{\text{AbsurdFalsePurpose}}$=0.24, 95% CI = [0.19, 0.29]). As expected, the highest proportion of suspicious participants were in the false purpose condition with the absurd experimental purpose that claimed to be about "Juggling Clowns." Those who believed they were deceived were also more confident in their belief than those who had not been deceived ($W_{one\text{-}sided}$=77,980 p<0.001). Yet, higher levels of suspicion are uncorrelated with the level of payoffs (rho=-0.04, p=0.203, *refer to Table S2*). Further, no interaction is found between conditions and level of suspicion in regression when predicting payoffs (*refer to Table S3*). In reviewing free text responses of what participants who had a suspicion of deception thought they had been deceived of (n=209), we find that the most common suspicion related to false information. A common suspicion among this subset was that researchers would not pay the bonuses as specified in the die roll task - a form of false purpose that could be considered for any future revisions of a deception taxonomy. This is despite generally high levels of self-reported trust in us to make necessary payments in this study ($M_{\text{TrustinUsforPayments}}$=4.3 (1: "Not at all", 5: "Very much"), 95% CI = [4.3, 4.4]) and for other researchers to do the same ($M_{\text{TrustinOthersforPayments}}$=4.3, 95% CI = [4.2, 4.3]). Nevertheless, this subset of MTurkers validate long-held concerns from economists that participant pools exposed to deception may not believe disclosures from researchers, thereby invalidating their ability to test hypotheses. The second most common suspicion reported related to false purpose.

Preferences and norms regarding false purpose

When asked whether researchers should be allowed to use false purpose deception, 78% of participants believed that it should be permitted on MTurk. Using the incentivized measure of social norms among MTurkers, we found, however, that participants expected their peers to be less tolerant towards false purpose deception. On average, there was an expectation that only 54% of other MTurkers would believe this type of deception should be permitted. The stronger endorsement of using false purpose when measuring individuals' personal views (rather than their (incentivized) expectations of peers) may reflect in part a reluctance to offend the researchers. The incentivized measure of peer expectations might therefore be a less biased measure of their beliefs, mitigating experimenter demand and social desirability bias(Bicchieri 2016).

Attention checks

In addition to measuring attention to stated experimental purpose, we measured time spent on the consent form and de-briefing. These did not vary across conditions ($H(3)_{Consent}$=2.1567, p=0.541, $H(3)_{Debriefing}$=6.62, p=0.085), nor with suspicion of being deceived in this experiment ($W_{two\text{-}sded}$= 99,753, p=0.849), suggesting that inattention or differential inattention does not drive these results.

Survey Control Variables

We found that most controls relevant to the experiment had no effect on the likelihood of reporting a winning outcome. With reference to the probit model (*refer to Table S2*), expectations that one was deceived in this or past experiments, ability to correctly identify the stated purpose of the experiment or the die roll had no impact on payoffs. As distinct from Study 1, MTurk experience had no impact on payoffs, though experience with die roll tasks did. The size of effect however was negligible (less than one tenth of a cent).

Demographic Control Variables

We find that only age has an influence on payoffs, and even then, it is minimal. For each additional year of age, the total payoff declines by less than one tenth of a percent (p<0.001, *refer to Table S2*).

Survey Findings

Experience of Deception

Deception appears to be commonly practiced on MTurk. Of our sample of 1,209 participants, 78% recalled being de-briefed regarding use of deception, while 11% did not recall being deceived and the remainder were unsure. ‘False purpose’ and ‘role deception’ were the most commonly identified forms of deception, with 65% and 66%, respectively, of those experiencing deception reporting having been exposed. ‘Unaware of measures’ was the least common type of reported deception, with 24% of previously deceived participants reporting that. Moreover, MTurkers commonly report being exposed to multiple types of deception (*see Figure 4b*). On average, participants had experienced four types of deception on MTurk.

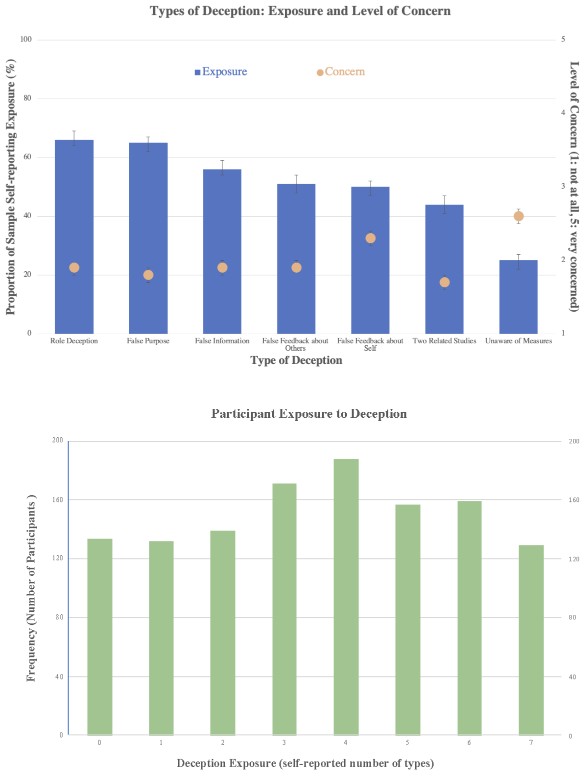


Fig. 4. a. Role deception and false purpose are the most common types of deception participants report having been exposed to in prior experiments, while ‘unaware of measures’ is the least common form. The levels of concerns are generally consistent at low levels across deception types with higher concerns noted with receiving false feedback about oneself and being unaware of measures in the experiment (e.g. filming, eye-tracking). **b.** The number of different types of deception participants

reported experiencing. In Study 2, we surveyed participants on seven different types of deception. Only 11% reported experiencing no deception on MTurk. On average, participants had experienced four different types of deception, while 11% reported experiencing all seven types of deception.

Attitudes towards Deception

While the exposure to deception appears to be common, the level of concern is bounded at low levels (see *Figure 4a*). There is some variation as would be expected (Kimmel 1998), but it was fairly limited and on average, participants are less than "moderately" concerned for all types of deception.

Spillover Effects on Trust in Science and Researchers

Consistent with the relatively low levels of concern regarding the use of deception, the majority of participants tended to expect relatively few adverse spillovers, with the exception of suspicion, from exposure to unspecified deception. *Refer to Figure 5 and Table 2.*

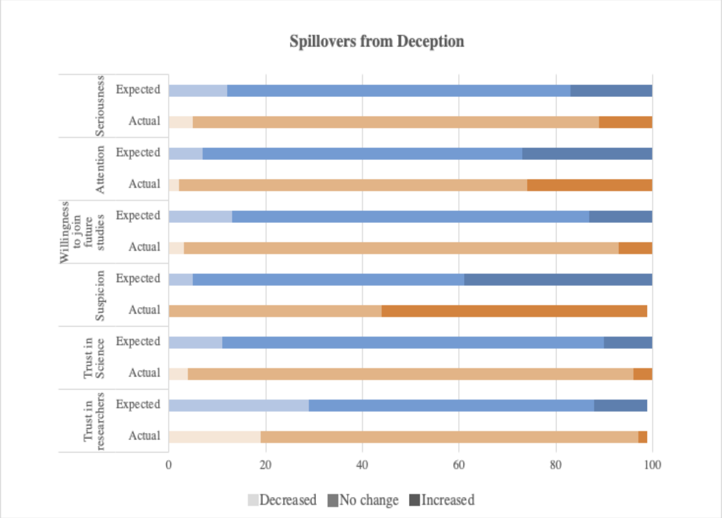


Fig. 5. The majority of participants expect no spillovers from deception on a range of measures. The one exception relates to suspicion, where those previously exposed to deception but not naive participants, reported increased suspicion. The figure shows the proportion of responses (decreased, no change, increased) for each type of spillover, split by whether a participant reported being previously exposed to deception. For 'naive' participants, the responses reflect expectations of spillovers from exposure to deception (n=248). For previously deceived participants (n=961), the responses reflect actual spillovers experienced post-deception.

Table 2: Spillovers from Deception on Future Behaviour

	Reported Prior Exposure to Deception – reporting actual effects (n=961) Proportion of Total (%)		**No Prior Exposure to Deception** – reporting anticipated effects (n=248) Proportion of Total (%)	
	No change	**Change**	**No change**	**Change**
Similar tasks	75 [72, 77]	25 [23, 28]	70 [64, 76]	30 [24, 36]
Different tasks	82 [80, 85]	18 [15, 20]	68 [62, 74]	32 [26, 38]

Note: 95% confidence intervals are presented in brackets.

One of the leading arguments against deception is that trust in researchers and in science in general (Kelman 1967; Baumrind 1979) can be undermined by deception. Krupat and Garonzik (1994) found survey evidence in support of this. Yet, (Epley and Huff 1998; Sharpe, G. Adair, and Roese 1992; S. S. Smith and Richardson 1983) do not find evidence of adverse effects of past deception on other researchers and on trust in psychological research. Our survey results do not point to an experienced or anticipated loss of trust in science but in researchers. This is especially pronounced among 'naive' participants (n=216): 29%, CI=[24%, 25%] of participants who reported not having previously experienced deception felt that their trust in researchers would be reduced after deception. In contrast, while 19% of previously deceived participants (n=961) report a loss in trust in researchers, for the remaining majority, the absence of such a loss may reflect their experienced deception not breaching the boundaries of the experimenter-participant contract (Epstein, Suedfeld, and Silverstein 1973).

Spillover Effects on (Stated) Future Behavior and Participation

Our survey results (*see Table 2*) suggest that there is a risk of provoking increased suspicion in future experiments. Epley and Huff (1998) find experimental support for this notion - even three months after the use of false feedback deception. Further, Krupat and Garonzik (1994) find survey evidence of this. Similarly, we find that the majority of previously deceived participants hold the view that past exposure to deception will provoke suspicion in future studies (55%, CI=(52%, 58%)), whereas naive participants hold weaker expectations of increased suspicion (39%, CI=(33, 45%)).

Despite the fairly common experience or expectation of increased suspicion after exposure to deception, this did not always transfer into reports of or expectations of changes in behavior. This is broadly consistent with our experimental findings reported above and with previous research (Krasnow, Howard, and Eisenbruch 2020), wherein increased suspicion of deception is not associated with a change in behavior. Of note, expectations of changed behavior in different future experimental tasks were more pronounced among the naive (M=32%, CI=[26%, 38%]) than the previously deceived population (M=18%, CI=[15%, 20%], $W_{one\text{-}sided}$=136,168, p<0.001).

This is in contrast to earlier experimental findings of actual behavior only changing with firsthand experience of deception, and not the possibility of deception (T. D. Cook and Perrin 1971; Christensen, n.d.). Still, consistent with expectations expressed by (Hertwig and Ortmann 2008a), we found that previously deceived participants more commonly report that deception might lead them to change their behavior in similar future tasks, relative to different future tasks ($W_{two\text{-}sided}$= 497,318, p-value <0.001). No differences between similar and different future tasks are predicted by naive participants ($W_{two\text{-}sided}$= 30,256, p= 0.651).

Previous exposure to deception has been argued to reduce the willingness for experimental subjects to participate in future experiments, creating selection biases. Jamison et al. (2008) find gender effects for selection attrition when using role deception; deceived women are less likely to return. On the other hand, there is survey evidence that deceived participants have a greater enjoyment of studies having been deceived (S. S. Smith and Richardson 1983) and expect others to enjoy such studies (Gerdes 1979) - which might suggest that such participants would be more willing to participate in the future. Here we find among both naive and previously deceived participants, that stated willingness to participate in future experiments is most commonly unchanged, with a leaning towards increased participation. Of note, no gender differences regarding willingness to participate in future experiments are identified among previously deceived ($W_{two\text{-}sided}$ = 112,220, p = 0.390) or deception-naive participants ($W_{two\text{-}sided}$ = 7,954, p = 0.535).

Spillover Effects on Survey Attention and Seriousness

We also assessed reported actual and expected spillovers from deception on participant seriousness and attention in future experiments. The vast majority of participants expected no change in their level of seriousness and attention after being exposed to or anticipating their response deception. Directionally, participants report it would in fact increase attention and seriousness in future experiments.

We are able to map these expectations with actual differences in behavior among our subjects. In contrast to the survey results, participants that reported having been previously deceived were more likely to be attentive than naive participants, according to three measures. Specifically, previously deceived participants were 13% more likely to correctly identify experimental purpose (*refer to Table S5*) and spent nearly double the time on the de-briefing page, measured in seconds ($M_{Deceived}$ = 13.4, CI = [12.3, 14.5], M_{Naive}= 7.2, CI = [5.8 , 8.6], $W_{one\text{-}sided}$=78,730, p<0.001). Deceived participants also spent more time on the consent form, but the size of the difference is not meaningful ($M_{Deceived}$ = 13.8, CI = [11.4, 16.3], M_{Naive}= 13.6, CI = [9.1, 18.1], W=99,628, p<0.001). However, these results may in part be driven by general learning and experience with research surveys and experiments among deceived participants who have on average ~5 months more experience on MTurk than naive participants (MTurk ($M_{Deceived}$ = 2.9, CI = [2.8, 3.1], M_{Naive}= 2.5, CI = [2.3 , 2.8], $W_{two\text{-}sided}$=109,667, p=0.024).

Discussion

Across two highly powered experiments, we do not find evidence that the nature of stated experimental purpose affects the measurement of honesty. In Study 1, we found that statements of true purpose, incomplete disclosure or a standard false purpose did not affect experimental outcomes in a 10-round coin flipping task. In Study 2, we conceptually replicated this finding with a one-shot die roll task, and added an additional condition, in which we test an “absurd” false purpose. We found that the majority of participants were attentive to the disclosed purpose (70%; 95% CI = [68%, 73%]) and that the absurd false purpose condition was effective in provoking suspicion of deception. Still, neither attentiveness nor suspicion were predictive of honesty.

Our findings are consistent with the idea of the ‘good subject’ (Orne 1962) or ‘faithful subject’ (Fillenbaum 1966), and confirmed by Spinner (1977): that is, even when participants believe they are being deceived, they follow the instructions carefully and execute tasks as if they had been deceived. Relatedly, false purpose may be considered legitimate by participants, thereby neutralising any effects from the suspicion of deception (T. D. Cook and Perrin 1971). A perceived negligible harm from our use of false purpose may also explain the lack of responsiveness to deception (C. P. Smith 1981) as it is not sufficient to break trust and trigger a retaliatory reaction against researchers (Epstein, Suedfeld, and Silverstein 1973). Certainly, participants’ common endorsement of and the low level of concern expressed regarding the use of false purpose in Study 2 would support this view. Still, the presumably more reliable incentivized measure of peer expectations suggested more reservations regarding experimenter use of false purpose. It is unclear whether this stems from concerns regarding harm, legitimacy or both.

This research has important theoretical and practical implications. Our findings suggest that results from these honesty tasks - of which 191 are reported in Gerlach et al. (2019) and 82 experiments in Abeler et al (2019) – are likely unbiased by the nature of stated purpose disclosed by the experimenter. We understand from both authors that, while not coded for, incomplete disclosure was the most common form for disclosing experimental purpose (Abeler 2021). We hope that our key findings help advance inter-disciplinary discussions in academia surrounding the use of deception, encourage reflections on how deception exposure can be better documented both on crowd-sourced platforms and laboratories, and spur further research on the effects—if any exist—of different deceptions on varying tasks.

We believe this research is relevant to work on replications, notably in the field. Variations in both IRB and field partner tolerances of deception can challenge efforts to conduct faithful replications where deception was involved in the original work. Even where an IRB is tolerant of deception usage, gatekeepers of participating institutions may be unwilling to permit the use of deception with their stakeholders. Our work may provide comfort - at least where honesty is being measured in commonly used tasks such as the coin-flipping and die roll tasks - that the use of incomplete disclosure instead of deception would not significantly affect results.

We also provide some insights regarding the apparent absence of past experimenters' work adversely spilling over to current experimentation. That is, we do not find evidence of (self-reported) past deception being associated with changed behavior in our studies. Finally, this revives work last conducted in the 1970s, with a new type of, and important, experimental participant—the crowdsourced worker—while also using large samples to enhance the robustness of the results. Given the widespread use of MTurk in social science research, a set of pre-registered studies with large sample sizes to answer these questions has been long overdue. We hope to contribute and revive this discussion to continue the best research practices across fields.

There, of course, are limitations to our findings. A key limitation relates to generalisability of the type of deception we studied. Participants are likely to vary in their responses to different types of deception and the variable harm and perceived legitimacy which comes with differing experimental designs. In this sense, 'false purpose' appears to be among the less egregious forms of experimenter deceptions so our findings may present the lower bound of the spillover effects that could emerge from other types of deception (e.g., false feedback) which appears to provoke more concern among our experimental participants (see survey results above).

Our work focuses on two of four common honesty tasks. Similar to Gerlach et al (2019), we do not find that honesty varies with deception in coin-flip and die roll tasks. However, this may not necessarily be the case with other honesty tasks (e.g. Sender-Receiver games which involve another participant in a strategic interaction) as their meta-analysis would suggest. Notably, their meta-analysis did not discriminate between types of deception. It may be that there is heterogeneity in the response to deception type (e.g. the existence of another participant that a participant is meant to interact with versus false purpose) and that this could also interact with the nature of the task.

The generalisability of our results are also limited by the nature of our sample. MTurk is one of many experimental platforms, and there are reasons to believe there are differences in participants between the platforms (Peer et al. 2017; Gupta, Rigotti, and Wilson 2021). With regard to honesty tasks in particular, Gerlach et al (2019) find that MTurkers do not engage in more dishonesty than economics students, non-economics students or ‘non-student’ participants in the die roll task, though substantially (21%) more cheating was found by MTurkers engaged in the coin flip task relative to economics students. Another popular online experimental pool to compare to our results in the future is Prolific Academic, which has gained considerable traction among academic researchers (Palan and Schitter 2018), emphasises the importance of honesty in on-boarding participants and has expressed an intolerance for dishonesty from participants (Wikely 2021). Also, our samples on MTurk, while large, were not recruited on a representative basis in light of budget considerations.

Results regarding the effects of deception may also vary over time. This is particularly true, given the high turnover of the MTurk population: 50% of the participant pool renews around every 7 months (Stewart et al. 2015). To ensure that our results are not driven by a single “wave” of participants on MTurk at any one time, we have mitigated the effects of high turnover by running experiments over an extended period of nearly four years.

Regarding our survey results, a key limitation is that we are measuring expected spillovers to changes in attitudes and behaviors in response to deception in general rather than actual behaviors. Our and others’ experimental results together suggest that the nature of deception is important in terms of consequences on participants. As such, our survey findings may be somewhat muted relative to the extremes that may emerge from both seemingly innocuous and highly concerning forms of deception.

Conclusion

Our work may not move researchers with some key concerns surrounding the use of deception. Indeed, a concern raised is that deception can permanently spoil the participant pool for future research: further work on actual behavioral spillovers in future studies from past exposure to deception is therefore needed. However, others might take some reassurance in that the use of false purpose deception is unlikely to bias behavior immediately in the study at hand (in our case, honesty behavior, as measured in two common honesty tasks). We hope that this work inspires further investigation of how different types of deception affect behavior in experiments and of spillovers from past deception.

Acknowledgements

We thank Larissa Samaan for coding assistance, Ralph Hertwig for helpful comments, and Erez Yoeli for reviewing the pre-registration for Study 1.

Transparency

Author Contributions

Z. Rahwan developed the study concept and created the surveys. All authors designed the study. Z. Rahwan analyzed and interpreted the data. Z. Rahwan drafted the manuscript. All authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

We gratefully acknowledge funding from Google ATAP and the London School of Economics.

Open Practices

All data and materials have been made publicly available via the OSF and can be accessed at https://osf.io/f6gmb/?view_only=2ad7305cce094ff4a349850dcbcc304e . The design and analysis plans for Studies 1 and 2 were pre-registered at AsPredicted (copies of the preregistration can be seen at the same OSF site).

References

Abeler, Johannes. Letter to Zoe Rahwan. 2021, August 13, 2021.

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. “Preferences for Truth-telling.” *Econometrica: Journal of the Econometric Society* 87 (4): 1115–53.

Adair, John G., Terrance W. Dushenko, and Rcl Lindsay. 1985. “Ethical Regulations and Their Impact on Research Practice.” *The American Psychologist* 40 (1): 59–72.

Advisory Committee on Human Radiation Experiments. 1996. *The Human Radiation Experiments*. Oxford University Press, USA.

“Altmetric – Business Culture and Dishonesty in the Banking Industry.” 2020. August 26, 2020. https://www.altmetric.com/details/2905465.

Ariely, Dan, and Michael I. Norton. 2007. “Psychology and Experimental Economics A Gap in Abstraction.” *Psychology and Experimental Economics* 16 (6): 336–39.

Barrera, Davide, and Brent Simpson. 2012. “Much Ado About Deception: Consequences of Deceiving Research Participants in the Social Sciences.” *Sociological Methods & Research* 41 (3): 383–413.

Baumrind, Diana. 1979. “IRBs and Social Science Research: The Costs of Deception.” *IRB* 1 (6): 1–4.

Bicchieri, Cristina. 2016. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Blackhart, Ginette C., Kelly E. Brown, Travis Clark, Donald L. Pierce, and Kelsye Shell. 2012. “Assessing the Adequacy of Postexperimental Inquiries in Deception Research and the Factors That Promote Participant Honesty.” *Behavior Research Methods* 44 (1): 24–40.

Bohannon, John. 2016. “Who’s Downloading Pirated Papers? Everyone.” *Science* 352 (6285): 508–12.

Bortolotti, L., M. Mameli, and Matteo Mameli. 2006. “Deception in Psychology: Moral Costs and Benefits of Unsought Self-Knowledge.” *Accountability in Research* 13 (3): 1–20.

Boynton, M. H., Portnoy, D. B., & Johnson, B. T. 2013. “Exploring the Ethics and Psychological Impact of Deception in Psychological Research.” *IRB* 35 (2).

Christensen, Larry. n.d. “The Negative Subject: Myth, Reality, or a Prior Experimental Experience Effect?” *Journal of Personality and Social Psychology* 35 (6): 392–400.

———. 1988. “Deception in Psychological Research: When Is Its Use Justified.” *Personality & Social Psychology Bulletin* 14 (4): 664–75.

Cohn, Alain, Ernst Fehr, and Michel Andre Marechal. 2014. “Business Culture and Dishonesty in the Banking Industry.” *Nature* 516 (729): 86–89.

Cook, Karen S., and Toshio Yamagishi. 2008. “A Defense of Deception on Scientific Grounds.” *Social Psychology Quarterly*. American Sociological Association. https://doi.org/10.1177/019027250807100303.

Cook, Thomas D., and Burton F. Perrin. 1971. “The Effects of Suspiciousness of Deception and the Perceived Legitimacy of Deception on Task Performance in an Attitude Change experiment1.” *Journal of Personality* 39 (2): 204–24.

Cooper, David J. 2014. “A Note on Deception in Economic Experiments.” *Journal of Wine Economics* 9 (2): 111–14.

Epley, Nicholas, and Chuck Huff. 1998. “Suspicion, Affective Response, and Educational Benefit as a Result of Deception in Psychology Research.” *Personality & Social Psychology Bulletin* 24 (7): 759–68.

Epstein, Yakov M., Peter Suedfeld, and Stanley J. Silverstein. 1973. “The Experimental Contract: Subjects’ Expectations of and Reactions to Some Behaviors of Experimenters.” *The American Psychologist* 28 (3): 212–21.

“Ethical Principles of Psychologists and Code of Conduct.” 2016. American Psychological Association. https://www.apa.org/ethics/code/ethics-code-2017.pdf.

Fillenbaum, Samuel. 1966. “Prior Deception and Subsequent Experimental Performance: The ‘Faithful’ Subject.” *Journal of Personality and Social Psychology*. https://doi.org/10.1037/h0023860.

Fischbacher, U., and F. Föllmi-Heusi. 2013. "Lies in Disguise—an Experimental Study on Cheating." *Journal of the European Economic Association* 11 (3): 525–47.

Gallo, Philip S., Shirley Smith, and Sandra Mumford. 1973. "Effects of Deceiving Subjects upon Experimental Results." *The Journal of Social Psychology* 89 (1): 99–107.

Gerdes, Eugenia Proctor. 1979. "College Students' Reactions to Social Psychological Experiments Involving Deception." *The Journal of Social Psychology* 107 (1): 99–110.

Gerlach, Philipp, Kinneret Teodorescu, and Ralph Hertwig. 2019. "The Truth about Lies: A Meta-Analysis on Dishonest Behavior." *Psychological Bulletin* 145 (1): 1–44.

Gino, Francesca, Maurice E. Schweitzer, Nicole L. Mead, and Dan Ariely. 2011. "Unable to Resist Temptation: How Self-Control Depletion Promotes Unethical Behavior." *Organizational Behavior and Human Decision Processes* 115 (2): 191–203.

Glinski, Richard J., Bernice C. Glinski, and Gerald T. Slatin. 1970. "Nonnaivety Contamination in Conformity Experiments: Sources, Effects, and Implications for Control." *Journal of Personality and Social Psychology* 16 (3): 478–85.

Gupta, Neeraja, Luca Rigotti, and Alistair Wilson. 2021. "The Experimenters' Dilemma: Inferential Preferences over Populations." *arXiv [econ.GN]*. arXiv. http://arxiv.org/abs/2107.05064.

Hertwig, Ralph, and Andreas Ortmann. 2008a. "Deception in Experiments: Revisiting the Arguments in Its Defense." *Ethics and Behavior* 18 (1): 59–92.

———. 2008b. "Deception in Social Psychological Experiments: Two Misconceptions and a Research Agenda." *Social Psychology Quarterly* 71 (3): 222–27.

Hey, John D. 1998. "Experimental Economics and Deception: A Comment." *Journal Of Economic Psychology* 19 (3): 397–401.

Huang, Karen, Regan M. Bernhard, Netta Barak-Corren, Max H. Bazerman, and Joshua D. Greene. 2021. "Veil-of-Ignorance Reasoning Mitigates Self-Serving Bias in Resource Allocation during the COVID-19 Crisis." *Judgment and Decision Making* 16 (1). http://journal.sjdm.org/20/201205/jdm201205.pdf.

Jamison, Julian, Dean Karlan, and Laura Schechter. 2008. "To Deceive or Not to Deceive: The Effect of Deception on Behavior in Future Laboratory Experiments." *Journal of Economic Behavior & Organization* 68 (3): 477–88.

Kajackaite, Agne, and Uri Gneezy. 2017. "Incentives and Cheating." *Games and Economic Behavior* 102 (March): 433–44.

Kelman, H. C. 1967. "Human Use of Human Subjects: The Problem of Deception in Social Psychological Experiments." *Psychological Bulletin* 67 (1): 1–11.

Kimmel, Allan J. 1998. "In Defense of Deception." *The American Psychologist* 53 (7): 803–5.

Krasnow, Max M., Rhea M. Howard, and Adar B. Eisenbruch. 2020. "The Importance of Being Honest? Evidence That Deception May Not Pollute Social Science Subject Pools after All." *Behavior Research Methods* 52 (3): 1175–88.

Krupat, Edward, and Ron Garonzik. 1994. "Subjects' Expectations and the Search for Alternatives to Deception in Social Psychology." *The British Journal of Social Psychology / the British Psychological Society* 33 (2): 211–22.

Mason, Winter, and Siddharth Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44: 1–23.

McDermott, Rose. 2013. "The Ten Commandments of Experiments." *PS - Political Science and Politics* 46 (3): 605–10.

Milgram, Stanley. 1963. "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology* 67 (4): 371 – undefined.

Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *The American Psychologist* 17 (11): 776.

Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17 (March): 22–27.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70 (May): 153–63.

Rahwan, Z., E. Yoeli, and B. Fasolo. 2019. “Heterogeneity in Banker Culture and Its Influence on Dishonesty.” *Nature* 575 (7782). https://doi.org/10.1038/s41586-019-1741-y.

Raymond, Nathaniel. 2019. “Safeguards for Human Studies Can’t Cope with Big Data.” *Nature*.

Roelcke, Volker. 2004. “Nazi Medicine and Research on Human Beings.” *The Lancet* 364 Suppl 1 (December): s6–7.

Roth, Alvin E. 2001. “Form and Function in Experimental Design.” *The Behavioral and Brain Sciences* 24 (3): 427–28.

Schuman, S. H., S. Olansky, E. Rivers, C. A. Smith, and D. S. Rambo. 1955. “Untreated Syphilis in the Male Negro; Background and Current Status of Patients in the Tuskegee Study.” *Journal of Chronic Diseases* 2 (5): 543–58.

Shalvi, Shaul, Jason Dana, Michel J. J. Handgraaf, and Carsten K. W. De Dreu. 2011. “Justified Ethicality: Observing Desired Counterfactuals Modifies Ethical Perceptions and Behavior.” *Organizational Behavior and Human Decision Processes* 115 (2): 181–90.

Sharpe, Donald, John G. Adair, and Neal J. Roese. 1992. “Twenty Years of Deception Research: A Decline in Subjects’ Trust?” *Personality & Social Psychology Bulletin* 18 (5): 585–90.

Sieber, Joan E., Rebecca Iannuzzo, and Beverly Rodriguez. 1995. “Deception Methods in Psychology: Have They Changed in 23 Years?” *Ethics & Behavior* 5 (1): 67–85.

Silverman, Irwin, Arthur D. Shulman, and David L. Wiesenthal. 1970. “Effects of Deceiving and Debriefing Psychological Subjects on Performance in Later Experiments.” *Journal of Personality and Social Psychology* 14 (3): 203–12.

Smith, Charles P. 1981. “How (Un)Acceptable Is Research Involving Deception?” *IRB: Ethics & Human Research* 3 (8): 1–4.

Smith, Stevens S., and Deborah Richardson. 1983. “Amelioration of Deception and Harm in Psychological Research: The Important Role of Debriefing.” *Journal of Personality and Social Psychology* 44: 1075–82.

Spinner, Barry, John G. Adair, and Gordon E. Barnes. 1977. “A Reexamination of the Faithful Subject Role.” *Journal of Experimental Social Psychology* 13 (6): 543–51.

Stewart, N., C. Ungemach, A. J. L. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, and J. Chandler. 2015. “The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers.” *Judgment and Decision Making* 10 (5): 13.

Stricker, L. J., S. Messick, and D. N. Jackson. 1967. “Suspicion of Deception: Implications for Conformity Research.” *Journal of Personality and Social Psychology* 5 (4): 379–89.

Taylor, Kevin M., and James A. Shepperd. 1996. “Probing Suspicion among Participants in Deception Research.” *The American Psychologist* 51 (8): 886–87.

“The Belmont Report.” 1979. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf.

Wikely, Steve. 2021. “New Prolific Features.” February 23, 2021. https://community.prolific.co/t/new-prolific-features/40/11.

Zimbardo, Philip G., Craig Haney, W. Curtis Banks, and David Jaffe. 1971. *The Stanford Prison Experiment*. Zimbardo, Incorporated.

Zizzo, Daniel John. 2010. “Experimenter Demand Effects in Economic Experiments.” *Experimental Economics* 13 (1): 75–98.

Supplementary Materials

Appendix 1

Original Definitions (Sieber (1995) as described by Hertwig & Ortman (2008)	**Revised Definitions, adapted to be meaningful from the perspective of an experimental participant**
False purpose. Participants are given, or cause to hold, false information about the main purpose of the study.	**False purpose:** Holding, or having cause to hold, false (not just incomplete) information about the main purpose of an experiment.
Bogus device. Participants are given false information concerning stimulus material.	**False information in stimulus materials.** Being presented with false information as part of the experiment. For example, you are given false answers of a potential dating partner, you are incorrectly informed that a human rather than an Artificial Intelligence (AI) generated text/image/sound/video.**
Role deception. Participants interact with participants about whose identity they have been given false information.	**Role deception.** Interacting with others in an experiment about whose identity you have been given false information. For example, another participant actually works in the research team, the researcher makes false claims about you being paired with a partner, including whether the partner is Artificially Intelligent (AI).***
False feedback regarding self. Participants are given false feedback about themselves.	**False feedback regarding self.** Receiving false feedback about yourself.
False feedback regarding others. Participants are given false feedback about another person.	**False feedback regarding others.** Receiving false feedback about others.
Two related studies. Two related studies are presented as unrelated	**Two related studies.** Two related studies or tasks are presented as unrelated
Unaware of measure. Participants are kept unaware that a study is in progress at the time of manipulation or measurement, or unaware of being measured (e.g. videotaped)	**Unaware of measures.** Being kept unaware that a study is in progress at the time of manipulation or measurement, or unaware of being measured (e.g. filming, eye-tracking). This excludes measures such as hidden click counts and question timers in on-line surveys.
Unaware of participation. Participants are kept unaware of being subjects in research.	(see above)

*** This is a common problem on MTurk (Krasnow et al, 2019, Summerville and Chartier (2013))**Bogus physical devices (e.g. lie-detector machines, Crutchfield apparatus) are of lesser concern in an on-line experimental environment.

Table S1: Study 1 Marginal Effects from Probit Models

Dependent Variable: Likelihood of Reporting a Winning Coin Toss	**Predictors**	**Model (1)**	**Model (2)** - with experimental controls	**Model (3)** - with demographics
Treatment (False Purpose omitted)	Incomplete Disclosure	0.02 [-0.02, 0.05]	0.02 [-0.01, 0.05]	0.02 [-0.01, 0.05]
	True Purpose	-0.02 [-0.05, 0.01]	-0.02 [-0.05, 0.01]	-0.02 [-0.05, 0.01]
	MTurk experience (years)		0.01* [0.00, 0.02]	0.01** [0.00, 0.02]
Expectation of Deception	Current Experiment (dummy)		0.03 [-0.01, 0.07]	0.03 [-0.01, 0.07]
	Past Experiment (dummy)		-0.02 [-0.05, 0.00]	-0.02 [-0.05, 0.01]
	Coinflip task experience (# tasks)		0.00 [-0.00, 0.00]	0.00 [-0.00, 0.00]
	Coinflip familiarity (dummy)		-0.01 [-0.05, 0.03]	-0.00 [-0.04, 0.03]
Last Coinflip	Over a year		-0.02 [-0.08, 0.05]	-0.02 [-0.08, 0.04]
(Within a year omitted)	Today		-0.04 [-0.18, 0.09]	-0.05 [-0.18, 0.08]
	Within last week		0.03 [-0.02, 0.09]	0.02 [-0.03, 0.08]
	Within last month		0.02 [-0.08, 0.11]	-0.00 [-0.10, 0.09]
	Conflip purpose known		-0.02 [-0.05, 0.01]	-0.02 [-0.05, 0.01]
	Age (years)			-0.00*** [-0.00, -0.00]
	Male			0.03* [0.01, 0.06]
	Higher Education			-0.02 [-0.05, 0.00]
Region	North East			-0.00 [-0.04, 0.04]
(Mid-West omitted)	Pacific			0.13 [-0.01, 0.27]
	South			-0.04* [-0.08, -0.01]
	West			-0.02 [-0.06, 0.02]
AIC		12,596	12,587	12,516
BIC		12,617	12,680	12,659
Log Likelihood		- 6,295	-6,281	- 6,238
Deviance		12,590	12,561	12,476
Num. obs.		9,270	9,270	9,250

Note: Responses for individuals are clustered and standard errors are robust. Figures in brackets denote 95% confidence intervals of the estimate. * p<0.05 ** p<0.01 *** p<0.001

Table S2: Study 2 Coefficients from Linear Regression Models

Dependent Variable: Payoff from die roll outcome (USD)	**Predictors**	**Model (1)**	**Model (2)** - with experimental controls	**Model (3)** - with demographics
	Intercept	33.40 *** (31.64 – 35.16)	35.60 *** (32.55 – 38.65)	39.00 *** (34.10 – 43.89)
Treatment (True Purpose omitted)	Incomplete Disclosure	0.89 (-1.58 – 3.35)	0.66 (-1.83 – 3.15)	0.58 (-1.90 – 3.06)
	False Purpose- Absurd	1.45 (-1.02 – 3.93)	1.57 (-0.95 – 4.09)	1.40 (-1.11 – 3.91)
	False Purpose - Standard	1.72 (-0.77 – 4.20)	1.56 (-0.94 – 4.06)	1.56 (-0.93 – 4.05)
	MTurk experience (years)		-0.19 (-0.59 – 0.21)	-0.08 (-0.48 – 0.33)
	Die roll experience (# tasks)		0.09 ** (0.02 – 0.15)	0.08 * (0.01 – 0.14)
	Suspicion of deception (dummy)		-1.81 (-4.16 – 0.54)	-2.09 (-4.43 – 0.25)
	Past experiment deception (dummy)		-1.32 (-3.55 – 0.90)	-0.72 (-2.97 – 1.52)
	Stated purpose identified (dummy)		-0.77 (-2.75 – 1.21)	-0.44 (-2.41 – 1.54)
	Die roll purpose identified (dummy)		-0.83 (-2.63 – 0.97)	-1.04 (-2.84 – 0.76)
	Age (years)			-0.13 *** (-0.20 – -0.05)
	Male			1.36 (-0.43 – 3.16)
	Higher Education			0.74 (-1.21 – 2.70)
	Relative Income			0.16 (-0.24 – 0.56)
	Political Preference (economic)			0.02 (-0.24 – 0.27)
	Religiosity			-0.38 (-0.81 – 0.05)
Observations		1,209	1,191	1,190
R^2		0.002	0.013	0.031

Figures in brackets denote 95% confidence intervals of the estimate. * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Table S3: Study 2 Linear Regression Results with interaction between treatment and suspicion

Dependent Variable: Payoff (USD)	**Model (1)**	**Model (2)**
Predictors	*Estimates*	*Estimates*
Intercept	33.68 ***	32.91 ***
	(33.68 – 33.68)	(32.91 – 32.91)
Incomplete Disclosure	1.08	1.13
	(1.08 – 1.08)	(1.13 – 1.13)
False Purpose - Absurd	1.26	1.10
	(1.26 – 1.26)	(1.10 – 1.10)
False Purpose - Standard	1.99	1.86
	(1.99 – 1.99)	(1.86 – 1.86)
Suspicion of Deception	-1.50	-1.00
	(-1.50 – -1.50)	(-1.00 – -1.00)
Incomplete Disclosure X	-2.45	-2.57
Suspicion of Deception	(-2.45 – -2.45)	(-2.57 – -2.57)
False Purpose - Absurd X	1.16	1.66
Suspicion of Deception	(1.16 – 1.16)	(1.66 – 1.66)
False Purpose - Standard X	-2.31	-2.13
Suspicion of Deception	(-2.31 – -2.31)	(-2.13 – -2.13)
Die Roll Experience (# tasks)		0.09 **
		(0.09 – 0.09)
Suspicion of Deception * Die Roll		-0.05
Experience (# tasks)		(-0.05 – -0.05)
Observations	1209	1191
R^2	0.006	0.011

Note: True Purpose variable omitted. Figures in brackets denote 95% confidence intervals of the estimate
* *p<0.05* ** *p<0.01* *** *p<0.001*

Table S4: Study 1 Marginal Effects from Probit Model of Suspicion

Dependent Variable: Suspicion of Deception in Current Experiment

Predictors	*Estimates*
Incomplete Disclosure	-0.03 (-0.08, 0.03)
True Purpose	0.00 (-0.06, 0.05)
Experience in Coin Flip tasks (#)	0.00 (-0.00, 0.01)
Deceived in past (dummy)	0.17*** (0.12, 0.21)
MTurk experience (years)	-0.01 (-0.03, 0.00)
AIC	805.51
BIC	834.5
Log Likelihood	-396.76
Deviance	793.51
Number of Observations	927

Note: False Purpose variable omitted. Figures in brackets denote 95% confidence intervals of the estimate ** p<0.05 ** p<0.01 *** p<0.001.*

Table S5: Study 2 Marginal Effects from Manipulation Check Probit Model

Dependent Variable: Correctly Identifying Stated Purpose	
Predictors	*Estimates*
Incomplete Disclosure	-0.08 (-0.15, -0.00)*
False Purpose - Absurd	0.17 (0.11, 0.23)***
False Purpose - Standard	0.03 (-0.04, 0.10)
Past Deception	0.13*** (0.06, 0.20)
AIC	1420.02
BIC	1445.51
Log Likelihood	-705.01
Deviance	1410.02
Number of Observations	1209

Note: True Purpose variable omitted. Figures in brackets denote 95% confidence intervals of the estimate * *p<0.05* ** *p<0.01* *** *p<0.001.*

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for Truth-telling." *Econometrica: Journal of the Econometric Society* 87 (4): 1115–53.

Akerlof George, A., and E. Kranton Rachel. 2000. "Economics and Identity." *The Quarterly Journal of Economics* 115 (3): 715–53.

Alexander, Anita, Michael Barnett-Cowan, Elizabeth Bartmess, Frank A. Bosco, Mark Brandt, Joshua Carp, Jesse J. Chandler, et al. 2012. "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science." *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 7 (6): 657–60.

"Altmetric – Business Culture and Dishonesty in the Banking Industry." 2020. August 26, 2020. https://www.altmetric.com/details/2905465.

"Altmetric – The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." 2020. August 27, 2020. https://sage.altmetric.com/details/29332953.

Aquino, Karl, and Reed Americus. 2002. "The Self-Importance of Moral Identity." *Journal of Personality and Social Psychology* 83 (6): 1423–40.

Ariely, Dan, and Michael I. Norton. 2007. "Psychology and Experimental Economics A Gap in Abstraction." *Psychology and Experimental Economics* 16 (6): 336–39.

Asch, Solomon E. 1955. "Opinions and Social Pressure." *Scientific American* 193 (5): 31–35.

Bargh, J. A., M. Chen, and L. Burrows. 1996. "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype-Activation on Action." *Journal of Personality and Social Psychology* 71 (2): 230–44.

Barrera, Davide, and Brent Simpson. 2012. "Much Ado About Deception: Consequences of Deceiving Research Participants in the Social Sciences." *Sociological Methods & Research* 41 (3): 383–413.

Becker, Gary S. 1968. *Crime and Punishment: An Economic Approach*. London: Palgrave.

Bortolotti, L., M. Mameli, and Matteo Mameli. 2006. "Deception in Psychology: Moral Costs and Benefits of Unsought Self-Knowledge." *Accountability in Research* 13 (3): 1–20.

Brunswik, E. 1955. "Representative Design and Probabilistic Theory in a Functional Psychology." *Psychological Review* 62 (3): 193–217.

Bryan, Christopher J., Elizabeth Tipton, and David S. Yeager. 2021. "Behavioural Science Is Unlikely to Change the World without a Heterogeneity Revolution." *Nature Human Behaviour* 5 (8): 980–89.

Camerer, C. F., A. Dreber, E. Forsell, T-H Ho, J. Huber, M. Johannesson, M. Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science." *Nature Human Behaviour* 2 (9): 637–44.

Cesario, Joseph. 2014. "Priming, Replication, and the Hardest Science." *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 9 (1): 40–48.

Chandler, Jesse, Cheskie Rosenzweig, Aaron J. Moss, Jonathan Robinson, and Leib Litman. 2019. "Online Panels in Social Science Research: Expanding Sampling Methods beyond Mechanical Turk." *Behavior Research Methods* 51 (5): 2022–38.

Cohen, Taya R., Scott T. Wolf, A. T. Panter, and Chester A. Insko. 2011. "Introducing the GASP Scale: A New Measure of Guilt and Shame Proneness." *Journal of Personality and Social Psychology* 100 (5): 947–66.

Cohn, Alain, Ernst Fehr, and Michel Andre Marechal. 2014. "Business Culture and Dishonesty in the Banking Industry." *Nature* 516 (729): 86–89.

Cohn, Alain, Michel André Maréchal, David Tannenbaum, and Christian Lukas Zünd. 2019. "Civic Honesty around the Globe." *Science*. https://doi.org/10.1126/science.aau8712.

Conrads, Julian, Bernd Irlenbusch, Rainer Michael Rilke, Anne Schielke, and Gari Walkowitz. 2014.

"Honesty in Tournaments." *Economics Letters* 123 (1): 90–93.

Cook, Karen S., and Toshio Yamagishi. 2008. "A Defense of Deception on Scientific Grounds." *Social Psychology Quarterly*. American Sociological Association. https://doi.org/10.1177/019027250807100303.

Cooper, David J. 2014. "A Note on Deception in Economic Experiments." *Journal of Wine Economics* 9 (2): 111–14.

Effron, Daniel A., Christopher J. Bryan, and J. Keith Murnighan. 2015. "Cheating at the End to Avoid Regret." *Journal of Personality and Social Psychology* 109 (3): 395–414.

Epstein, Yakov M., Peter Suedfeld, and Stanley J. Silverstein. 1973. "The Experimental Contract: Subjects' Expectations of and Reactions to Some Behaviors of Experimenters." *The American Psychologist* 28 (3): 212–21.

"Ethical Principles of Psychologists and Code of Conduct." 2016. American Psychological Association. https://www.apa.org/ethics/code/ethics-code-2017.pdf.

Fillenbaum, Samuel. 1966. "Prior Deception and Subsequent Experimental Performance: The 'Faithful' Subject." *Journal of Personality and Social Psychology*. https://doi.org/10.1037/h0023860.

Fischbacher, U., and F. Föllmi-Heusi. 2013. "Lies in Disguise—an Experimental Study on Cheating." *Journal of the European Economic Association* 11 (3): 525–47.

Gächter, Simon, and Jonathan F. Schulz. 2016. "Intrinsic Honesty and the Prevalence of Rule Violations across Societies." *Nature*. https://doi.org/10.1038/nature17160.

Gallo, Philip S., Shirley Smith, and Sandra Mumford. 1973. "Effects of Deceiving Subjects upon Experimental Results." *The Journal of Social Psychology* 89 (1): 99–107.

Gerlach, Philipp, Kinneret Teodorescu, and Ralph Hertwig. 2019. "The Truth about Lies: A Meta-Analysis on Dishonest Behavior." *Psychological Bulletin* 145 (1): 1–44.

Gino, Francesca, Shahar Ayal, and Dan Ariely. 2013. "Self-Serving Altruism? The Lure of Unethical Actions That Benefit Others." *Journal of Economic Behavior & Organization* 93 (September): 285–92.

Gino, Francesca, Maurice E. Schweitzer, Nicole L. Mead, and Dan Ariely. 2011. "Unable to Resist Temptation: How Self-Control Depletion Promotes Unethical Behavior." *Organizational Behavior and Human Decision Processes* 115 (2): 191–203.

Glinski, Richard J., Bernice C. Glinski, and Gerald T. Slatin. 1970. "Nonnaivety Contamination in Conformity Experiments: Sources, Effects, and Implications for Control." *Journal of Personality and Social Psychology* 16 (3): 478–85.

Gneezy, Uri, Agne Kajackaite, and Joel Sobel. 2018. "Lying Aversion and the Size of the Lie." *The American Economic Review* 108 (2): 419–53.

Gupta, Neeraja, Luca Rigotti, and Alistair Wilson. 2021. "The Experimenters' Dilemma: Inferential Preferences over Populations." *arXiv [econ.GN]*. arXiv. http://arxiv.org/abs/2107.05064.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *The Behavioral and Brain Sciences* 33 (2-3): 61–83.

Hertwig, Ralph, and Till Grüne-Yanoff. 2017. "Nudging and Boosting: Steering or Empowering Good Decisions." *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 12 (6): 973–86.

Hertwig, Ralph, and Andreas Ortmann. 2008. "Deception in Experiments: Revisiting the Arguments in Its Defense." *Ethics and Behavior* 18 (1): 59–92.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124.

———. 2008. "Why Most Discovered True Associations Are Inflated." *Epidemiology* 19 (5): 640–48.

Jamison, Julian, Dean Karlan, and Laura Schechter. 2008. "To Deceive or Not to Deceive: The Effect of Deception on Behavior in Future Laboratory Experiments." *Journal of Economic Behavior & Organization* 68 (3): 477–88.

Kajackaite, Agne, and Uri Gneezy. 2017. "Incentives and Cheating." *Games and Economic Behavior* 102 (March): 433–44.

Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams, Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating Variation in Replicability

Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–90.

Kouchaki, Maryam, and Francesca Gino. 2016. "Memories of Unethical Actions Become Obfuscated over Time." *Proceedings of the National Academy of Sciences* 113 (22): 6166–71.

Krasnow, Max M., Rhea M. Howard, and Adar B. Eisenbruch. 2020. "The Importance of Being Honest? Evidence That Deception May Not Pollute Social Science Subject Pools after All." *Behavior Research Methods* 52 (3): 1175–88.

Kristal, Ariella S., Ashley V. Whillans, Max H. Bazerman, Francesca Gino, Lisa L. Shu, Nina Mazar, and Dan Ariely. 2020. "Signing at the Beginning versus at the End Does Not Decrease Dishonesty." *Proceedings of the National Academy of Sciences of the United States of America* 117 (13): 7103–7.

Kuhn, Thomas. 2021. "The Structure of Scientific Revolutions." In *Philosophy after Darwin*, 176–77. Princeton University Press.

Lane, David M., and William P. Dunlap. 1978. "Estimating Effect Size: Bias Resulting from the Significance Criterion in Editorial Decisions." *The British Journal of Mathematical and Statistical Psychology* 31 (2): 107–12.

Leif, Nelson, Uri Simonsohn, and Joe Simmons. 2021. "Evidence of Fraud in an Influential Field Experiment About Dishonesty." *Data Colada* (blog). August 17, 2021. http://datacolada.org/98.

Mason, Winter, and Siddharth Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44: 1–23.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." *JMR, Journal of Marketing Research* 45 (6): 633–44.

McDermott, Rose. 2013. "The Ten Commandments of Experiments." *PS - Political Science and Politics* 46 (3): 605–10.

Milgram, Stanley. 1963. "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology* 67 (4): 371 – undefined.

Milgram, Stanley, Leon Mann, and Susan Harter. 1965. "The Lost-Letter Technique: A Tool of Social Research." *Public Opinion Quarterly* 29 (3): 437.

Miller, Dale T., and Daniel A. Effron. 2010. "Psychological License. When It Is Needed and How It Functions." In *Advances in Experimental Social Psychology*, 43:115–55. Academic Press Inc.

Mohan, G. 2014. "Banking Industry Culture Primes for Cheating, Study Suggests." *LA Times*, November 21, 2014. http://www.latimes.com/science/sciencenow/la-sci-sn-cheating-bankers-20141119-story.html.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (January): 0021.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 359 (6251). https://doi.org/10.1126/science.aac4716.

Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *The American Psychologist* 17 (11): 776.

Pashler, H., and Eric Jan Wagenmakers. 2012. "Introduction to the Special Section on Replicability." *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 7 (6): 528–30.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70 (May): 153–63.

Rahwan, Zoe, Oliver P. Hauser, Ewa Kochanowska, and Barbara Fasolo. 2018. "High Stakes: A Little More Cheating, a Lot Less Charity." *Journal of Economic Behavior & Organization* 152 (August): 276–95.

Rosenbaum, Stephen Mark, Stephan Billinger, and Nils Stieglitz. 2014. "Let's Be Honest: A Review of Experimental Evidence of Honesty and Truth-Telling." *Journal Of Economic Psychology* 45 (December): 181–96.

Serra-Garcia, Marta, and Uri Gneezy. 2021. "Nonreplicable Publications Are Cited More than

Replicable Ones." *Science Advances* 7 (21). https://doi.org/10.1126/sciadv.abd1705.

Sieber, Joan E., Rebecca Iannuzzo, and Beverly Rodriguez. 1995. "Deception Methods in Psychology: Have They Changed in 23 Years?" *Ethics & Behavior* 5 (1): 67–85.

Smith, Charles P. 1981. "How (Un)Acceptable Is Research Involving Deception?" *IRB: Ethics & Human Research* 3 (8): 1–4.

Spinner, Barry, John G. Adair, and Gordon E. Barnes. 1977. "A Reexamination of the Faithful Subject Role." *Journal of Experimental Social Psychology* 13 (6): 543–51.

Stewart, N., C. Ungemach, A. J. L. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, and J. Chandler. 2015. "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers." *Judgment and Decision Making* 10 (5): 13.

Stricker, L. J., S. Messick, and D. N. Jackson. 1967. "Suspicion of Deception: Implications for Conformity Research." *Journal of Personality and Social Psychology* 5 (4): 379–89.

Taylor, Kevin M., and James A. Shepperd. 1996. "Probing Suspicion among Participants in Deception Research." *The American Psychologist* 51 (8): 886–87.

Yong, Ed. 2012. "Replication Studies: Bad Copy." *Nature* 485 (7398): 298–300.

Zhong, Chen-Bo, Katie A. Liljenquist, and Daylian M. Cain. 2009. "Moral Self-Regulation : Licensing and Compensation." *Psychological Perspectives on Ethical Behavior and Decision Making*, 75–89.