ALL THE WORLD'S A STAGE:

ADAPTIVE IMPRESSION MANAGEMENT AND PROSOCIALITY

Mia Karabegović

Submitted to:

Central European University

Department of Cognitive Science

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

Primary supervisor: Christophe Heintz

Secondary supervisor: Dan Sperber

Budapest, Hungary

2021

Declaration of authorship

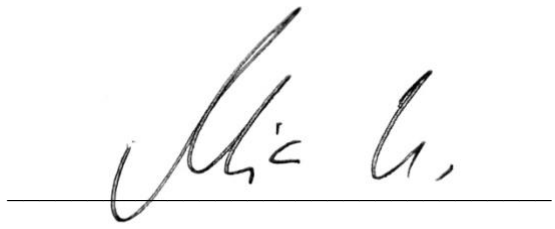
I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or which have been accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgment is made in the form of bibliographical reference.

The present thesis includes work that appears in the following papers/manuscripts:

Heintz, C., **Karabegovic, M.**, & Molnar, A. (2016). The co-evolution of honesty and strategic vigilance. *Frontiers in psychology, 7*, 1503.

Other parts of the thesis will be submitted for publication with the following co-authors:

Chapter 1: Christophe Heintz
Chapter 2: Christophe Heintz
Chapter 3: Amanda Rotella and Pat Barclay
Chapter 4: Anand Murugesan, Mahendran Chokkalingam and Christophe Heintz
Chapters 5 and 6: Anand Murugesan and Christophe Heintz



Mia Karabegović

Abstract

Audience effects – the differences which emerge due to one's belief about whether or not their behaviour will be observed – have been documented in a variety of seemingly unrelated phenomena. In the present work, we first propose a unifying perspective of this effect based on evolutionary theory - in Chapter 1, we outline the benefits of an evolved impression management mechanism, its relevant inputs and cues thereof, the ways in which environmental features could influence its outputs (impression management strategies), and review of a number of findings we believe fit under the umbrella of this functional interpretation. In subsequent chapters, we empirically investigate the sensitivity to observation qualified by subtle changes in audience features. Chapter 2 investigates whether observers' awareness of one's knowledge about strategic incentives influences prosocial choice and subsequent observers' evaluations of actors' prosocial behaviour in more or less strategic contexts. Chapter 3 addresses the importance of future benefits one can expect from an audience in decisions to make an initial prosocial choice to advertise their trustworthiness and, on the flipside, how observers' trust is modulated by the knowledge about these strategic incentives and choices. Chapters 4 and 5 shift the focus from cooperation to rule abidance: in Chapter 4, we address the relation between self-framing effects and local attitudes about social rules using two different methods (coordination games and self-report surveys); in Chapter 5 we investigate the influence of rule origins and leader intentions; while Chapter 6 presents an experiment about assortment and cost of rule following on rule abidance. We summarize the results we find in support for the idea of a fine-grained impression management mechanism sensitive to audience features in prosocial contexts in Chapter 7.

Acknowledgments

If someone told me they'd known exactly what a PhD would entail before they started it, I'd be their sceptical audience: with so many individual and circumstantial differences, the idea of this journey being predictable is unfathomable to me. Despite the common (and the not-so-common) hurdles on the way, my own term as a PhD student overshot my expectations in ways I couldn't have foreseen. This is mostly due to my primary supervisor, Christophe Heintz: saying his support has been essential to finishing this thesis seems trite (even if true). I've been lucky to experience genuine acknowledgment while working with Christophe, which I know from previous experience, academic and otherwise, isn't easy to find. Christophe – thank you for treating me as an equal since we met; for not sparing me the difficult questions; for (usually) being convinced by my answers. For making this process *fun*. For your understanding and flexibility. I hope you're pleased with this product of our cooperative venture. My sincere thanks also goes to Dan Sperber, my second supervisor, for his kindness; for always being ready with incredibly astute remarks on-the-go, conveyed in a laughably easy manner.

I am grateful to Anand Murugesan from CEU's School of Public Policy for his readiness to collaborate across disciplinary boundaries and the valuable insights and efforts in making parts of this thesis happen. To Pat Barclay for being a willing, able and available advisor during my visit to the University of Guelph, whose approach to evolutionary psychology remains an inspiration. To Igor Mikloušić and Josip Burušić, for providing the first opportunities to try my hand in research. I'd like to thank the CEU for the financial support. The Department of Cognitive Science and the lovely folks of the SOA group for their feedback throughout the years. Eszter Salamon, Vanda Derzsi and Nejra Rizvanovic – for their help in making these

experiments happen. Györgyné Finta – for making *everything* happen. My participants, online and offline, across the world.

On a more personal note, I'd like to thank my mother Julijana and father Ognjen, for their unconditional acceptance. Everything I've done right in my life is a reflection of the kind of parents you've been. My grandparents, who imparted lessons one can't learn from peer-reviewed journals: Mira, for showing me the value of generosity; Vida, for teaching me about the importance of rationality; Eduard, for his relentless support and trust in my abilities. My uncle Boris and aunt Azra – for providing cherished comic relief in the chaos that was often our day-to-day. To Danilo – my one-man IT crowd. We've been unlucky in some regards, but we have certainly been happy in our unique way – thank you all, for making lemons into lemonade.

To my friends from forever-ago: Vedrana, who has been my True North for twenty-four (!) years and counting; and Lukrecija, who has the superpower of making train-hopping across Europe as enjoyable as memorizing statistical formulae. Thank you both for being my anchors.

There are people I met during the PhD who have made the last seven years *a life.* Sam and Shannon, who are as much fun in Budapest, Toronto, Sarajevo and on Zoom. Nazlı and Luke, who never failed to provide much-needed laughter. Thomas and Martin, who kept the Jungle Office alive. Gina, a great write-up commiserate. Nirvana, who was always up for spending an afternoon in the many gardens of Erzsébetváros, for spicing those up with jokes in our native language. Amanda who is, legend has it, my research twin and has been an immense joy to run study ideas by and work with. Aurora, whose friendship was the unexpected silver lining during the pandemic, who never fails to brighten my day with her emails and the promise of future adventures (and baked goods).

I especially want to thank two women without whom I couldn't imagine this PhD. Oana, the best friend I could've hoped to meet on my very first day at CEU: thanks for the Netflix suggestions, fashion advice, random talks at 4 AM, weather reports, podcast recommendations, help with statistics and R. Most of all, though – thanks for being that bit of home away from home I didn't even think I would miss. Helena, my Monday-at-Kuplung companion, willing participant to random photography projects, science heroine, favourite reader-of-things-I-shouldn't-be-writing; who is always there to encourage my craziest ideas, as only the best of friends are. Thanks for being that rare person with whom the laws of Math do not apply – even when we are both deep in the negative, our addition always makes shared evenings into positives instead. That's some gift – one I will always cherish.

Finally, I want to thank Petar who probably deserves a chapter, not a paragraph, in this story. Thanks for the songs. The intricate dinners. For your support in all things I dare to imagine; for imagining those I don't yourself and pushing me in their direction. For coming along on the ride and being ridiculously easy to live with, hundreds of miles or an arm's length away. For the last eight years in which you've been my proof-reader, therapist, guinea pig, cheerleader, one-man band, chef, and most importantly – almost from the moment we met – my best friend. Here's looking at you, kid!

(And because I have some space left, I'd like to thank Dan Bejar a.k.a Destroyer for 'Tinseltown Swimming in Blood', that kept me awake through many nights I've spent writing and is a worthy anthem for this thesis.)

Table of Contents

Introduction

Modern humans live in an environment marked by apparent contradictions, populated as it is with trolls and bots; strangers hiding behind innocuous JPEGs waiting to ruin reputations on account of social *faux pas* (Ronson, 2015). One can always become a monster in their own right, too – clogging shared spaces with selfies and frequent updates of one's daily life, leading to significant academic interest in the links of social media use and personality traits such as narcissism, selfishness and empathy (e.g. Alloway, Runac, Quershi, & Kemp, 2014; Casale & Banchi, 2020). Academic and public interest in these topics is contrasted to articles about unexpected instances of selflessness or helping, which periodically appear online and spread 'virally' as evidence of humans' prevailing kindness and generosity. These stories have topics that stretch from unlikely do-gooders who find and return prized possessions (going against a self-interested 'finders keepers' philosophy), wide-ranging, cultural phenomena like the Neapolitan tradition of "suspended coffee" – a coffee you can pay for in advance for somebody who can't afford it (Pianigiani, 2014); or curious pay-it-forward incidents, like the one when a sequence of 167 drivers at a US McDonald's drive-through all paid for the orders of the people in the car behind them (Kindelan, 2017). The common features which might contribute to the success of such stories are that they provide information humans are likely 'wired' to find interesting, i.e. they map onto evolved cognitive mechanisms (for an example from another domain, see Miton & Mercier, 2015), and at the same time be sceptical about.

The title of this dissertation, taken from Shakespeare's As You Like It, calls into mind the sociological work of Erving Goffman, The Presentation of Self in Everyday Life (1959), in which he was among the first to provide arguments for the importance of impression

management (and, in fact, use the term) for understanding behaviour. Goffman (1959) likened social interactions to theatrical performances performed by 'actors', emphasizing definitions of situations and the influence individuals exert to change or maintain them in order to project the impressions they wish others to have of them. He provided a detailed framework of the many parts of this 'performance' which serve towards the desired impression, positing that human behaviour is often determined by others' expectations, which are in turn derived from social norms regarding different social roles and mutually agreed-upon definitions of the context in which it is performed.

While this thesis has been inspired in more direct ways by work in evolutionary psychology of cooperation and biological markets theory (Noë & Hammerstein, 1995; Barclay, 2016), the aspects first recognized by Goffman have remained relevant – the experiments presented in the following chapters tackle both the factors which modulate the desire to manage impressions depending on context (Chapters 2, 3 and 4), as well as the influence of local rules and others' expectations on prosocial and impression management-related choices (Chapters 4, 5 and 6). The abovementioned scepticism about self-presentation and curiosity about feel-good instances of prosociality thus make up two sides of the same coin (or the same thesis), connected by selection pressures related to partner choice ecologies and resulting in a cognitive mechanism of impression management on the one hand, which has the function of benefiting the actor by increasing their chances for mutualistic interactions, and a 'sceptical' mechanism of strategic vigilance on the other, which serves to weed out self-interested impression managers and is geared towards picking the best available partners for future cooperative ventures.

The focus of our investigation are so-called audience effects, i.e. differences between private and public behaviour. Broadly conceived as behaviours which change depending on

observation, they have been documented in several domains in psychological literature, which can be roughly separated into three main categories: (i) performance effects – e.g. social facilitation (Triplett, 1898; Bond & Titus, 1983), stereotype threat (Zanna & Pack, 1975); (ii) effects on prosocial and antisocial behaviour – e.g. contributions in economic games (Hoffman, McCabe, & Smith, 1996), moral licensing (Rotella, 2020), disinhibited behaviour in deindividuation (Diener, 1977); and (iii) expressed beliefs and attitudes – e.g. conformity and compliance (Bond, 2005; Sowden, et al., 2018), attitude change (Malkis, Kalle, & Tedeschi, 1982), self-enhancing biases (Sedikides, Herbst, Hardin, & Dardis, 2002), or explicit stereotyping (Plant & Devine, 1998). In Chapter 1, we discuss the way in which audience effects found in these social psychology phenomena can be re-conceptualized as outputs of an evolved mechanism for impression (reputation) management, and the potential benefits of this perspective for future research directions.

Another dimension of distinguishing audience effects relates to the type of experimental manipulation used to achieve the perception of observation or anonymity, as well as who the observers are and how much information about them the actors are privy to. The simplest variations include comparisons between choices made privately and choices made in front of an audience (e.g. Rege & Telle, 2004). Other studies have varied levels of identifiability (for instance, with photographs displayed to observers or not, Andreoni & Petrie, 2004) or levels of the observability of the action itself (for instance, whether a church donation is made into a closed or open donation box, Soetevent, 2005). Minimal cues that one is being observed can also nudge prosocial choice. Haley and Fessler (2005), for instance, showed the effect that the mere presence of drawn eyes can have on behaviour in economic games. This ‘eye-cue’ effect has been replicated in both lab (Burnham & Hare, 2007; Rigdon, Ishii, Watabe, & Kitayama, 2007)

and field contexts (Bateson, Nettle, & Roberts, 2006; Ernest-Jones, Nettle, & Bateson, 2011). Studies have also documented the importance of social norms (Bateson, Callow, Holmes, Roche, & Nettle, 2013; Kawamura & Kusumi, 2017) and the length of exposure to the 'false' cue (Sparks & Barclay, 2013), among others, as relevant features which modulate the target behaviours. Finally, changing features of the observer also impact behaviour. For instance, people will behave differently if they believe they are being observed by in-group and out-group members (Malkis, et al., 1982; Mifune, Hashimoto, & Yamagishi, 2010). Importantly, what actors believe their audience believes can also affect their behaviour (Alexander and Weil, 1969; Andreoni & Bernheim, 2009). Chapters 2 and 3 address some of the gaps in the cooperation literature related to the impact of audience features and beliefs on prosocial behaviour, while Chapters 4 and 5 deal with the influence of social rules on impression management strategies and behaviour in economic experiments, respectively.

The main hypothesis driving this thesis is that actors adaptively manage their (costly) prosocial impressions by taking into account the relevant audience features which reflect on their potential payoffs from future interactions with its members and the value of a given strategy in generating the desired impression of oneself in the observers (presenting oneself as a valuable co-operator). The flipside hypothesis refers to the audience members themselves and posits they will attend to the same contextual cues as impression managers in order to form predictions about future behaviour, using not only the outcome of observed (prosocial) actions to guide these predictions, but also the (plausible) underlying intentions.

In Chapter 1, we provide an elaboration of the idea of impression management as a set of adaptive cognitive mechanisms with audience effects as one of its outcomes and look at a variety of social psychology phenomena through this lens, arguing for their origins in impression (and

reputation) management. We also outline a theoretical model of hypothesized influences of observers' individual differences, as well as intermediate concepts and exogenous factors which are likely to influence competition on biological markets and consequently, impression management strategies. In Chapter 2, we test the hypothesis that the road to effectual impression management is 'paved with covert intentions'. Specifically, we focus on the audience's knowledge of actors' strategic incentives, and how this knowledge interacts with the decisions to signal prosociality, either overtly or covertly, in a partner-choice paradigm. We also explore evaluations of agents performing various prosocial actions in contexts which differ with regard to audience value and relationship to the recipient of the helping action to show that 'publicness' is not enough to lead to a self-defeating effect in signalling prosociality. Instead, attributions are fine-tuned to the variables relevant for the actor's future benefits.

Chapter 3 is also focused on audience features, specifically – audience quality and the expected benefits from interacting with its members. In an online economic experiment, we look at decisions to signal generosity to observers with randomly assigned competence levels (which lead to different expected benefits). In the same study, we also address observers' decisions to trust actors who had different information about the quality of their audience. Chapters 4, 5 and 6 address social rule abidance. In Chapter 4, we investigate framing effects and self-enhancement biases in ethnographically relevant scenarios in India, comparing two different methods of eliciting ethically sensitive responses related to social rules and the influence of social acceptability of the allegedly unethical choices on reports about one's (hypothetical) willingness to engage in said unethical behaviours. In Chapter 5, we examine the influence of rule origins, while Chapter 6 looks at the effects of rule costliness and its importance for subsequent assortment on abidance.

Chapter 1: The adaptive value of impression management

Audience (or observer) effects can be defined as a "change in behaviour caused by being observed by another person, or the belief that one is being observed by another person" (Hamilton & Lind, 2016). In this sense, the psychological phenomena we've previously mentioned qualify as audience effects: they show a sensitivity to the presence of others when performing a certain action. The notion of the audience effects we examine in this chapter is closely related to self-presentation and impression management, the latter describing "any behaviour by a person that has the purpose of controlling or manipulating the attributions and impressions formed of that person by others" (Tedeschi & Riess, 1981, p.3).[1]

While the concept of impression management has lately found most traction in organizational psychology, with a focus on its relevance in professional contexts (Bourdage, Roulin, & Levashina, 2017), it was previously used by social psychologists who questioned 'intrapsychic' (among other, cognitive) explanations of phenomena like cognitive dissonance, attitude change, conformity and compliance and argued for 'contextualist' approaches (Tetlock & Manstead, 1985; see also Leary, Raimi, Jongman-Sereno, & Diebels, 2015). These authors argued for a social account of the these phenomena – for example, in the case of cognitive dissonance, Tedeschi, Schlenker and Bonoma (1971) proposed an explanation that went beyond a variety of then-established theories which relied on the need for (internal) cognitive consistency (which would result in attitude change in forced compliance paradigms) and posited that this intrinsic need, in actuality, reflects a social concern to appear consistent to others in

[1] Throughout this thesis, we use the more general term of impression management as opposed to self-presentation, as other authors have done before (e.g. Tedeschi & Riess, 1981; Leary & Kowalski, 1990).

order to retain credibility and rationalize (i.e. justify) the counter-normative behaviour. Jellison (1981) went a step further with an even more behaviourism-flavoured approach to attitudes, claiming that social approval was one of the key motivations for phenomena such as conformity and consistency in behaviour, and stressing the influence of the material consequences of social approval and immediate situational variables.

Psychological explanations in this tradition were largely directed through a lens of social desirability: the myriad aforementioned effects were thus conceived as a consequence of the motivation to be valued by one's social circle and to comply with the norms they deem relevant (some of which are universal, like consistency). While this might be true for some categories we mentioned in the introduction (most notably, effects on prosocial behaviour and expressed beliefs and attitudes), other instances of audience effects need not be consciously geared towards making a good impression, even if they have the same function, i.e. that of reputation management. For example, audiences can direct one's attention and induce gaze following (Milgram, Bickman, & Berkowitz, 1969), which is in and of itself an audience effect, but more likely the result of informational rather than reputational concerns. The bystander effect (for a more recent review, see Fischer et al., 2011) is perhaps the most obvious example of an audience effect, however, the belief that one is observed in this case is also unlikely to be crucial, or something the actor is directly aware of – instead, the influence of the audience is actualized in the information it provides about the nature of the potential emergency and the need for action (Latané & Darley, 1968).

In most other instances, however, impression management likely plays a role in producing audience effects, via dedicated cognitive mechanisms which direct behaviour towards socially desirable ends. We believe this to be the case because, on the whole, it is beneficial for

humans to live in communities which appreciate their exhibited qualities and regard them as potentially valuable partners. In our view, impression management, as a psychological concept with its long research tradition, could gain more traction if characterized as a set of mechanisms that have the evolved function of managing others' beliefs about oneself. Crucially, we see impression management as hinging on evolved cognitive capacities rather than deliberate intentions to make a positive impression, avoiding some of the negative connotations which have been pinned to the concept. Furthermore, using a functional, evolutionary characterization of impression management can help unify explanations of various phenomena which show sensitivity to observation, while teasing apart those audience effects which are distinct and don't have reputation management as their primary function.

1.1. Impression management and biological markets

Psychology as a science had mostly been concerned with proximal causation (the 'how' of things) before the inception of the evolutionary psychology research program in the 1990s (Barkow, Cosmides, & Tooby, 1992). Evolutionary explanations have since gained traction by demonstrating the importance of considering psychological mechanisms from a functional perspective, especially with regards to elucidating the relevant features of input stimuli which activate said mechanisms and their dedicated mental processes. Taking a functional perspective can bring us closer to answering chicken-or-egg dilemmas, and affords more fine-grained predictions about the circumstances in which certain outputs - like audience effects - will occur.

Impression management can and does come about through a variety of proximal mechanisms – such as the 'warm glow' one feels after performing a generous action (Andreoni, 1990), or a preference for behaving in the most socially desirable manner the situation allows

(Alexander & Rudd, 1981). While these and similar mechanisms can contend for explaining how certain behaviours (like prosociality) come about, they fall short in telling us why they do in the first place. Our definition of impression management invokes the evolutionary function of behavioural changes under observation: we do not contend that it is always achieved through deliberate processes in which actors have the explicit intention of making favourable impressions (though it can be). However, we do predict that psychological mechanisms dedicated to impression management will take, as inputs, those stimuli on which evolutionary success is likely to be based and produce effects similar to what would be expected from an intentional 'gaming' of impressions.

In order to address this multi-realisability of the ultimate function of impression management (through the intentional management of impressions – even 'faking' – desirable qualities; and audience effects that come about 'automatically'), we can borrow an example from the mating market. Various human dispositions can be distilled to the underlying function of having children - but this outcome can be achieved through different proximal mechanisms. One of them is a conscious decision to have children: the intention with the content which is exactly the same as the ultimate function of the mechanism. However, humans are also motivated by sexual desire, the content of which is not the intention to produce offspring; while its function is still the same (it reliably had this effect without modern affordances such as birth control). Similarly, through proximal mechanisms of impression management which generate intentions and motivations such as wanting to portray oneself as cooperative or an otherwise desirable partner, or through feeling good when one acts prosocially, one can also achieve the same biological function of the mechanism – securing long-term partnerships that can aid in one's survival as well as reproductive success.

What, then, is the function of impression management? To answer this question, we turn to research from the sphere of the evolution of cooperation, namely, by taking into account partner choice ecology and biological markets theory (BMT; Noë & Hammerstein, 1994, 1995), which makes use of (economic) market analogies to provide a blueprint of the various features of partner-choice environments during which (proximate) mechanisms have evolved. BMT is based on the idea of market-like competition for desirable partners – those who can provide one with the most benefits in subsequent interactions. Furthermore, it takes into consideration that most people will prefer to cooperate with those high in partner value, which leads to assortative matching (Barclay, 2016); as well as extending indirect reciprocity models by drawing attention to the relative levels of cooperation on the market and broadening the definition of what makes one a valuable partner as depending on both the supply and demand of certain traits in the environment, and including traits beyond a mere propensity to reciprocate (Barclay, 2013).

BMT assumes a natural variation in the supply and demand of particular traits (or 'commodities') – the existence of individual differences in the relevant traits, which can influence performance costs and reflect on the strategies used to compete for partners, for instance in the sphere of generosity (Barclay & Reeve, 2012). As others have noted, this type of selection is comparable, or rather, super-ordinate to processes like sexual selection (Barclay, 2016). In fact, studies in the latter domain have also shown that market-like dynamics operate in the relationship between mate choice and mate preference (e.g. Wincenciak et al., 2015). A change in the supply and demand ratio of specific qualities, as well as the influence of exogenous factors (e.g. depletion of game, technological innovations), can make certain skills more or less essential and shift competition to other domains.

Given that evolutionary success hinges on one's survival as well as reproduction, the ability to discern valuable (cooperative or sexual) partners from cheaters or impostors emerges as one of the more relevant tasks our ancestors would have had to solve in the environment of evolutionary adaptedness (EEA; Tooby & Cosmides, 1992). On the other hand, the ability to manage impressions and consequently reputations also becomes an indispensable skill for competition in both domains. Acting prosocially – benefiting others at a cost without any immediate goal of securing rewards for the self – can thus be seen as an instance of 'competitive altruism' (Roberts, 1998; Van Vugt & Hardy, 2006) or an impression management strategy devised to improve (or hold constant) one's image score or reputation in a market of cooperators and secure more beneficial opportunities for future interactions with desirable partners, as well as outside options should a particular current partnership fail.

1.1.1. How impression management enhances fitness

If we conceive of audience effects as outcomes of an impression management mechanism which has the function of influencing others' perceptions of oneself in order to gain access to fitness-enhancing future partnerships, it becomes clear that the presence of observers alone is not enough to lead to what we observe in the literature. There are several steps at which psychological mechanisms dedicated to impression management should take inputs from the environment (see Figure 1.1.) and at which modulating factors could influence the motivation to manage impressions. These are the causal chains through which impression management leads to benefits on a market of cooperators, which only start with observation.

The observer (or the audience) first needs to be aware of the actor doing x, which is a behaviour indicative of a certain underlying disposition (X) and the targeted impression (e.g.

helping a tourist at a railway station pay for a tram ticket to get to their hotel because they don't have the right currency could be indicative of the helper's prosociality).[2] In this stage, it is crucial for the audience to attend to the behaviour or its outcome and perceive it as having been carried out by the actor. In the next step, the audience should update their beliefs about the actor's disposition X, which implies they should either already have a representation of the actor and their reputation in X, or 'fill in the blanks'. This doesn't necessarily mean that observers have to link one's face or name with the action (though being visible or seeing one's partner is certainly a reliable cue of being identifiable): what matters is the dependency on the identifier for future interactions. (For example, partner-matching in economic games or one's avatar and username in online communities are what impressions or reputations can be otherwise assigned to.) According to these updated beliefs, the willingness of audience members to select the actor as a partner in future mutualistic opportunities can be upgraded, downgraded, or stay the same.[3] In the cases where it's upgraded, when such an opportunity arises, the observers might pick the actor as a partner, leading to a mutually beneficial interaction.

The factors embedded in these causal chains, such as identifiability, the affordance for belief updating and the effect of action x on the belief about disposition X, should influence the actor when deciding whether or not to perform x, i.e. they should modulate audience effects. This conceptualization is in many ways similar to the crucial factors preceding impression management as described by Leary and Kowalski (1990): (i) goal relevance – where the

[2] The types of behaviours which will be most affected by impression management, in our view, are those of high importance for one's inclusive fitness (e.g. behaviours related to advertising traits relevant to mate selection and cooperative partner value), and especially those behaviours where the inference x → X is not direct, i.e. where the behaviour which leads to the impression is not always a reliable sign of the trait.

[3] Belief updating and the resultant willingness to interact with the actor crucially also depends on the credibility of action x as an indicator of disposition X. We address this more specifically in section 5, 'The Catch-22 of Prosocial Impression Management'.

publicness of the action is of central importance, as well as one's dependency on the target of the self-presentation and the expected probability of future contact with them; (ii) the value of desired goals – the scarcity or abundance of what is to be achieved by projecting the desired impression; and (iii) the discrepancy between one's desired and current image.

In the next section, we provide the basis of an evolutionary psychology-inspired research program for impression management by analysing the variables which are likely to modulate audience effects, along with the relevant cues and ultimate, super-ordinate factors these variables and cues are meant to provide information about. In order to do so, we use insights from partner choice and biological markets theory to form hypotheses and draft a comprehensive framework from which audience effects can be considered.

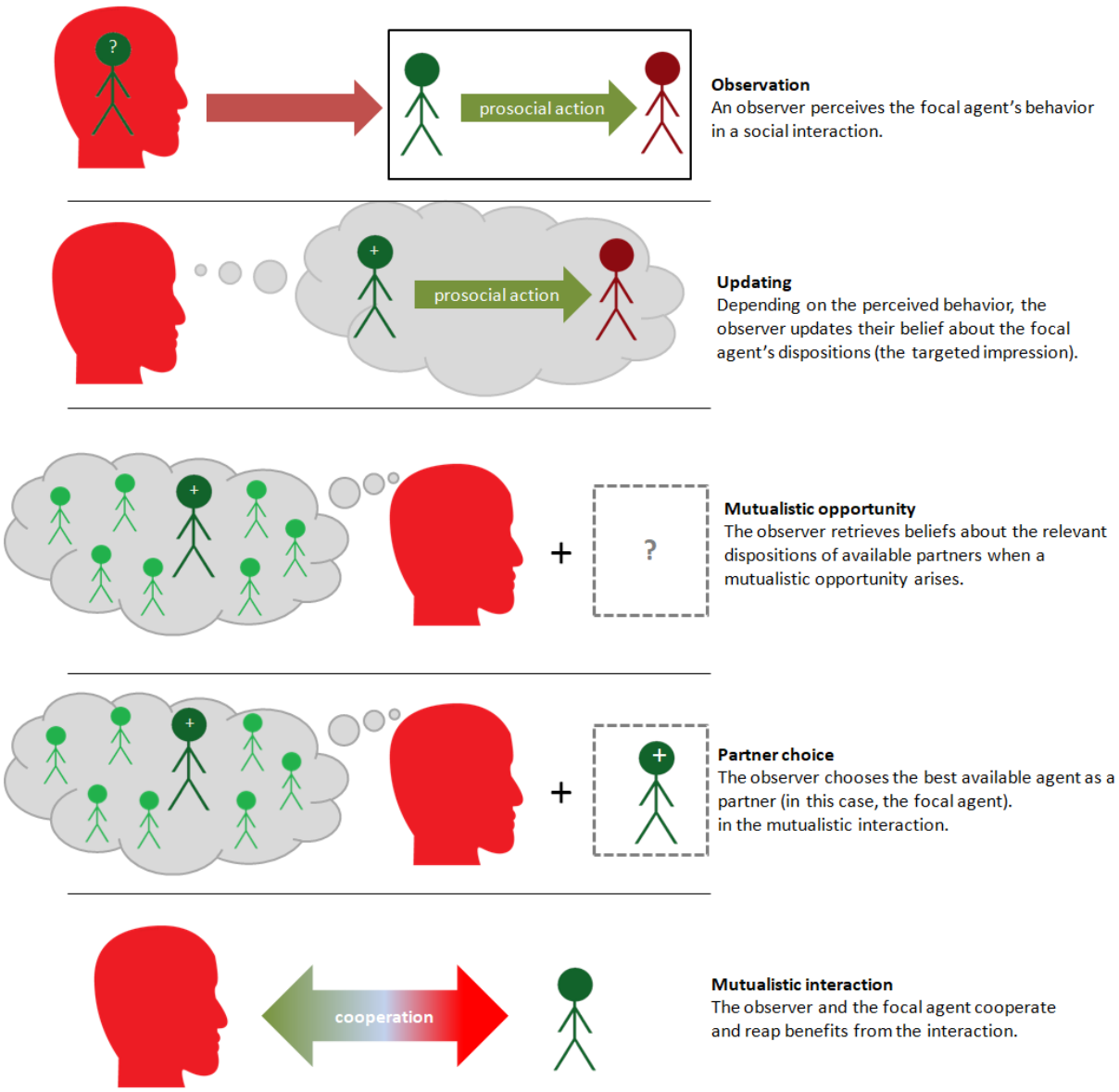


Figure 1.1. The cognitive causal chains through which impression management mechanisms achieve fitness-related benefits.

1.2. The ABCs of impression management and audience effects

Who are the partners with whom it is best to cooperate? Barclay (2013; 2016) identifies three relevant dimensions for selecting partners: their ability to confer benefits, their willingness to confer/share those benefits, and their availability to do so. Consequently, observers who have one or more of these characteristics are those who are most valuable to impress - they're the partners one would want to select for future cooperative interactions, and be selected by them in return.

Our functional conceptualization of impression management integrates these attributes into higher-level factors which, we posit, are the ultimate drivers of audience effects and the motivation to manage impressions of partner value. Specifically, we incorporate them into factors of the probability of future interactions and the perceived advantageousness or anticipated benefits from these interactions. We further provide a comprehensive framework of the related intermediate variables, as well as cues which should be attended to and modulate the outputs of impression management mechanisms.

1.2.1. When is impression management worth the cost?

An actor's choice of behaviour which serves impression management motives should depend on the expected benefit of said choice (including its costs) outweighing the expected benefit of not performing the behaviour in question. We define a benefit as any increase in an individual's inclusive fitness, which might come about both through securing valuable cooperative partnerships and opportunities on the mating market. While we focus mostly on the former aspect, they are often difficult to disentangle. For instance, studies have shown that prosociality can also have an effect on the desirability of long-term romantic partners

(Ehlebracht, Stavrova, Fetchenhauer, & Farrelly, 2018) and that generosity is increased in men who think their decisions will be observed by attractive females (Raihani & Smith, 2015).

What kind of costs and benefits can one encounter in situations where impression management is likely? Firstly, as we point to above, acting in a certain manner can increase one's opportunities for valuable future partnerships while failing to do so can lead to missing out on these opportunities. Furthermore, if one not merely fails to help, but also behaves selfishly or cheats in observable social interactions, one risks being punished (Fehr & Gächter, 2000) or ostracized (Feinberg, Willer, & Schultz, 2014). These outcomes are fairly straightforward, but there are other, more subtle, consequences which can influence one's social payoffs from impression management, too.

For instance, in situations of low credibility of a prosocial choice (in front of audiences likely to make an attribution of strategic motives instead of prosocial ones), the benefit of *not* making the prosocial choice might be preserving one's reputation as a consistent social actor, if not a generous one (Andrews, 2001). Else, when one already has a reputation of being a good partner, the benefits of additional advertising in ambiguous situations might be small, but the benefits of not bringing scepticism about one's intentions into the mix might outweigh these potential benefits through the loss of credibility that could be incurred by impression management. Similarly, in moral licensing, having previously established an impression of prosociality can make it seem more feasible for one to cheat 'a little' in future interactions, if one's reputation (or moral self-concept) is unlikely to be damaged (for a general overview of the effect, see Merritt, Effron, & Monin, 2010). This seems to be especially true when the first reputation-enhancing action has been observed and when the actions are morally ambiguous (Rotella, 2020).

Given these constraints – some of which are seldom apparent – how are the cost-benefit computations underlying impression management done for them to result in adaptive outcomes? In our view, impression management mechanisms depend on two main features related to the audience: the probability of future interactions with its members and the expected benefits of the observers (audience members) updating their beliefs about the actor, increasing their willingness to choose them as a partner in mutualistic interactions.[4] While this rendering seems simple at first glance, one might be reasonably sceptical that humans can perform such calculations or are equipped to gauge the many different factors. How does one know the probability of interacting with a person – especially one they have never met before – in the future? By the same token, how can one correctly determine the benefit such an interaction might confer? It should be noted we do not presume that either of the proposed underlying factors are evaluated directly. Rather, the information inherent to different situations which afford impression management, one's prior knowledge of an audience via reputation, or other characteristics of audience members can serve as intermediate factors which are funnelled through the respective cognitive mechanisms and result in one's choices.

A clarification can again be found in an example from the mating market, where the underlying function of partner preferences is increasing one's reproductive success. While this can be achieved through a number of long- or short-term strategies, depending on one's value as a potential mate among others, let us take the most obvious example of reproductive value, which refers to the (average) expected number of future offspring, i.e. how many children one is likely to contribute to the next generation from a given point in their life span (Buss & Schmitt,

[4] These benefits can also be indirect (other than interactions with a specific observer). For instance, observers who are well-connected in one's social group or market can influence one's reputation and future opportunities in an exponential manner through gossip (Wu, Balliet, & Van Lange, 2016).

1993). This was an especially important adaptive problem for men choosing their partners, as women are more constrained by biological factors in the number of children they can bear. How do men determine reproductive value? After all, it is not something that can be directly assessed, similar to the factors we propose as relevant audience characteristics. Instead, one can rely on intermediate factors such as health or age to approximate reproductive value. Neither health nor age, however, are readily obvious (unless one has access to an individual's birth certificate or full medical check-up report), so men can in turn attend to cues such as physical appearance (facial symmetry, skin, etc.) or observable behaviour (Buss & Schmitt, 1993). Because these cues have been shown to be more reliable than not – because the men attending to them had more reproductive success over time – preferences can reasonably have developed for a certain class of cues (a preference as to what is found physically attractive).

In the same vein, we posit that humans have developed a sensitivity to the evidence or cues of certain intermediate factors which have been reliably correlated with the probability of future interactions and their expected benefits. It seems trivial to assert that we use a wide range of external and contextual cues in the absence of first-hand knowledge about the potential partner's previous behaviour or reputation. Less obviously, we hypothesize that the mechanisms dedicated to partner choice – and subsequently impression management tailored to a partner choice market – are much more flexible than distributive preferences (such as 'be fair'), and that decisions about partnering-up are influenced by fine-grained cognitive mechanisms, instead of relying solely on observation cues. This should especially be true when actors have access to explicit information which contradicts the information that was crucial to the development of a cue, i.e. why it was originally attended to (its evolutionary function).

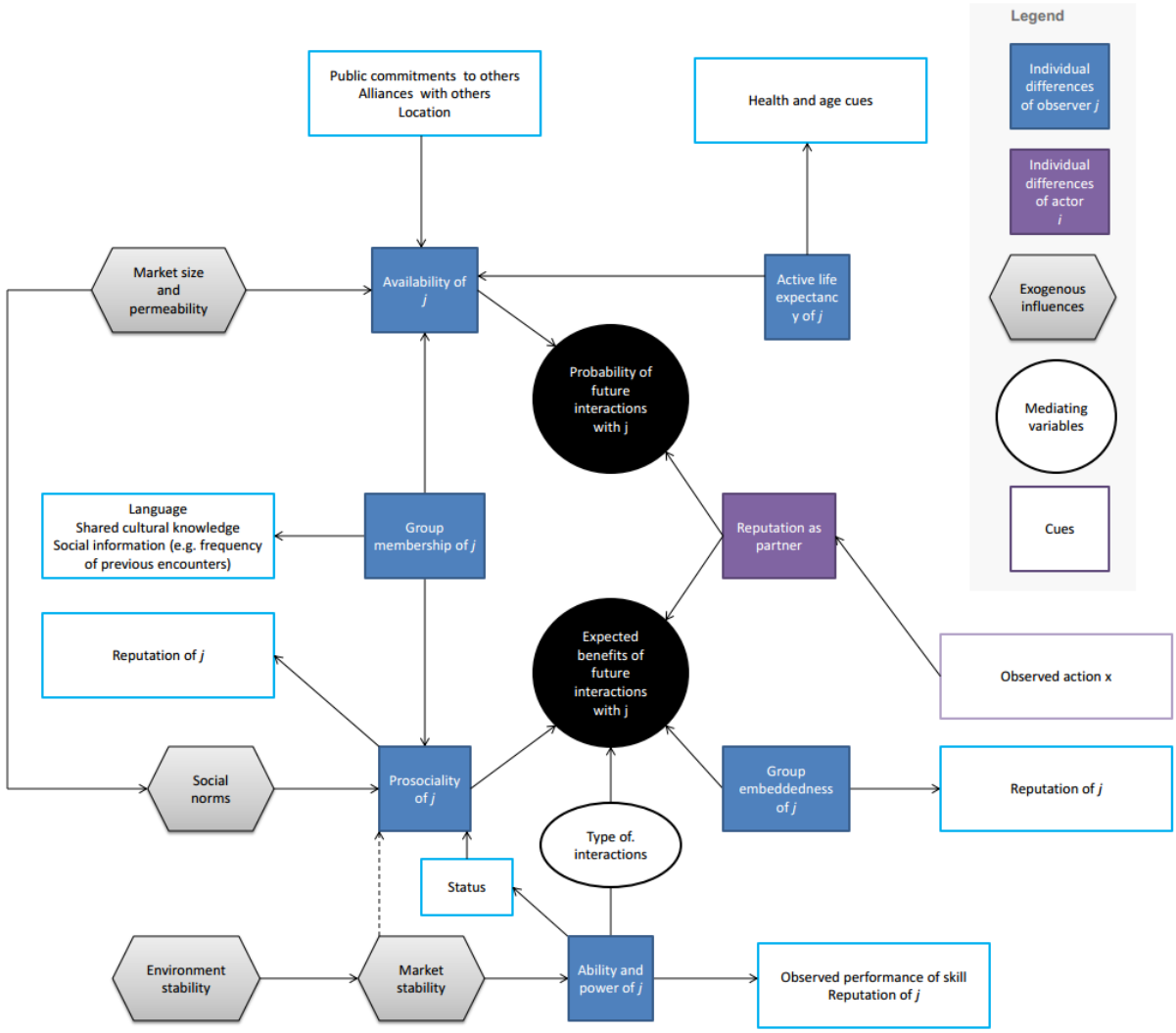


Figure 1.2. A theoretical model of individual actor and audience characteristics which influence impression management, including exogenous factors, intermediate variables, and cues thereof.

Additionally, certain cues can serve as indicators of more than one factor. For instance, group membership can be a cue of both the probability of future interactions and the willingness to confer benefits, if the observer is in the same group as the actor, through a process such as generalized reciprocity (Yamagishi & Kiyonari, 2000). Health cues, like physical fitness, are an indicator of the probability of interactions (inasmuch as they make it likelier one is going to be around to provide benefits at a future point in time), the number of expected interactions (for the

same reason), as well as the expected benefits themselves (if one is sickly, it is unlikely they will be able to consistently provide any form of benefits). To get a better grasp of these relationships, we summarize the main variables and their connections in Figure 1.2.

1.2.2. Audience size and the number of expected interactions

Before delving into the individual-level variables related to observers and their influence on the benefits of impression management, we need to address one aspect of the audience as a whole - the number of interactions the actor can expect with its members. Audience size is the most obvious cue of this factor, if the benefit is taken as a consequence of a single observed action. The bigger the audience, the more potential partners and potential subsequent interactions, and the more motivation to manage impressions (if other variables are held constant, i.e. if the probability of interactions and the expected benefits are the same).

However, the *plausible* number of interactions one can have with audience members is also important. If one 'performs' for an audience of ten thousand, it is improbable they will have interactions with all the observers. Similar to other group phenomena (e.g. conformity; Bond, 2005), there is likely a threshold at which audience size plateaus in terms of calculating future benefits or altering behaviour, which is constrained by the size of groups in which the mechanisms evolved. Human ancestors in the EEA did not have platforms such as twitter or Facebook to broadcast self-enhancing signals to tens or hundreds of thousands audience members at once. What this also means is that humans – at least a decade ago – were ill-equipped to tailor their impression and reputation management strategies to large, heterogeneous audiences. Social blunders and misjudging the size of one's audience can have a detrimental effect on one's image, and the type of reputational and material costs that can follow are real: as

in the case of Justine Saccho who was publicly shamed, fired and ostracized over a tasteless joke on Twitter in 2013 (Ronson, 2015).

While narratives about social media often take on negative connotations in how the technology is used and question people's understanding of its repercussions (harkening back and forth between different variations of the mismatch hypothesis; see Li, van Vugt, & Colarelli, 2018), studies have shown that social media users are learning to incorporate the information about their audiences into their impression management strategies, regardless of the potential counter-intuitiveness of having a wide reach. Particularly, the self-presentation of social media users seems to be geared toward the "strongest" audiences – those with the highest value to the user and the strictest standards combined (Marder, Joinson, Shankar, & Thirlaway, 2016) – which is likely the most adaptive strategy in this context.

Furthermore, the number of encounters one expects with each audience member should also impact the willingness to manage impressions, as it increases the number of beneficial interactions one stands to gain or lose from, even more so than the mere number of observers in the audience. Imagine giving two presentations to an audience of fifty colleagues: one in the very beginning of a five-year contract at the workplace, and one in the fourth year, knowing you will be moving on to another university soon. While the audience features and size don't change, and neither does your immediate probability of interacting with its members tomorrow or in a month, you might feel more motivated to give a good talk in the first case than in the second as the

'shadow of the future'[5] is longer, and you stand to lose on many more instances of mutualistic opportunities due to the time constraints.

Finally, a single behavioural choice likely has a 'shelf-life' which is as long as the time to the next (observed) interaction that affords impression management, because reputations – relevant as they are – should be constantly updated. This shelf-life is likely tied to the frequency of one's interactions with the audience. On the one hand, the expected frequency of interactions might be expected to influence impression management in a way similar to the probability of future interactions – to linearly increase its likelihood. However, as the 'shelf-life' of one instance of impression management is also longer with audiences one is infrequently interacting with, and the fact that there are less opportunities to engage in reputational correction and/or less possibilities to manage impressions, this effect could also go in the other direction. We expect that this variable should interact with audience value, such that the highest levels of impression management would be observed in the cases of high audience value and infrequent impression management opportunities, as opposed to when one has frequent opportunities to manage one's impressions with either a high- or low- value audience.

Though availability is often overlooked in research in favour of other relevant variables, it is interesting to note that some of its constraints have been – for the most part – removed in modern society, where cooperation can and does take place globally and oftentimes more quickly. These new affordances come with risks of their own – such as the likelihood of being deceived and reliance of reputation mechanisms which aren't necessarily trustworthy (Tennie,

[5] The shadow of the future is most often defined in terms of anticipated future interactions. Economic experiments with finite and infinitely repeated games investigating the impact of the 'shadow of the future' show that the probability of future interactions significantly increases cooperation (Bó, 2005), and that the level of cooperation grows in fixed-partner scenarios in infinitely repeated games (Duffy & Ochs, 2009).

Frith, & Frith, 2010). While these dangers certainly exist, it is interesting to note how technology and culture have 'outsourced' the relevant cognitive tasks to similar embodied versions of reputation tracking mechanisms such as user ratings and review aggregators (Heintz, 2006). The speed with which these services emerged can also be used as an argument for the importance of the information it provides - namely, of 'gossip' about partners one is considering cooperating with.

1.2.3. Probability of future interactions

As Leary and Kowalski (1990) note, the probability of future interactions is a factor which influences the motivation to manage impressions - studies which investigated modulation of self-enhancement biases provide one general line of evidence about the importance of anticipating future interactions (e.g. Wortman, Costanzo, & Witt, 1973; Sedikides et al., 2002), research on the shadow of the future in economic experiments provides another (Bó, 2005). Though tied to the number of expected interactions mentioned previously, the higher-order factor of the probability of interacting with members of an audience is different in that it can be more directly influenced by one instance of observed prosocial choice.

The probability of future interactions can mostly be gauged through cues of availability, some of which are straightforward. Physical inaccessibility and time constraints would both have been strong cues that one will not be able to cooperate with a given individual in the future (in other words, one needed to be in the same place, at the same time - and with few commitments to others). Active life expectancy, indicated by health and age cues, is also a good cue of

availability: if *j* is dead or sick, *i* is unlikely to cooperate with them in the future.[6] Other cues of availability include information about one's (public) commitments to others and alliances, as these can reduce future need for partners. If an audience member *j* is already committed to cooperating with a large enough number of others, and finds these relationships beneficial, it is less probable they will be looking for outside options, given the aforementioned physical and time constraints as well as the inherent risk and potential cost of partner-switching. On the other hand, in trying to keep their outside options open, people tend to keep their rankings of friends and alliance information as close to the chest as possible (DeScioli & Kurzban, 2009), which makes this information more difficult to gauge.

Group membership is another candidate which influences availability and subsequent probability of future interactions. Cues like language or shared cultural traditions – in other words, features which can serve as proxies of group membership – alert one to the fact that an observer is a member of one's group and as such would be more available for future interactions, by virtue of being in the same geographical location or social circle. The experiments showing even minimal groups based on trivial characteristics affect reasoning and behaviour about in- and out-group members point to the fact that it was likely an important cue (e.g. Tajfel, et al., 1971), with additional evidence from developmental psychology in the sensitivity to the types of cues which stand in for group membership and influence in-group favoritism, like accents do in 5-year olds (e.g. Kinzler, Shutts, DeJesus, & Spelke, 2009), or minimal group cues in even younger children (Richter, Over, & Dunham, 2016).

[6] Note, however, that while this might be the case when forming new partnerships for mutualistic interactions, the relationships and altruism towards one's kin (Ashton, Paunonen, Hermes, Jackon, 1998) and long-term friends (DeScioli & Kurzban, 2011) are likely distinct, and based on different relational models (Fiske, 1992).

Finally, the reputation of the actor *i* as a partner also influences the probability of future interactions. Those who previously make observed prosocial choices are selected more often for future interactions in economic games (Barclay & Willer, 2007; Sylwester & Roberts, 2010), and incurring intentional costs to benefit one's group leads to increased evaluations of partner desirability, as well as increased attributions of altruism (Delton and Robertson, 2012) .

1.2.4. Expected benefits

Several intermediate audience features affect the benefits an actor stands to gain from interactions with its members: ability and power (mediated by the types of interactions with agent *j*), prosociality, group embeddedness, and group membership (through prosociality and generalized reciprocity).

Ability is closely related to cues of competence and possession of valuable skills or qualities – these can range from displaying athletic and hunting prowess in small-scale societies (Bird, Smith, & Bird, 2001) to conspicuous consumption reflecting status or access to resources (i.e. wealth) in more modern settings (Sundie et al., 2011; Boone, 1998). As such, abilities, power and status are closely linked, especially in the case of the latter two which are often hard to disentangle. In our model, we use a definition of power which is primarily focused on an agent's control over others' outcomes, and position status as a potential cue of social power, given its likely influence on resource access and formation of moral values (of what is considered desirable behaviour), while acknowledging that status on its own doesn't necessarily lead to the same type of social influence as power (Fiske & Berdahl, 2007). In and of itself, high status has been shown to decrease prosociality in both adults and children (Guinote, Cotzia, Sandhu, & Siwa, 2015), while conversely prosociality can also help one *acquire* status (Kafashan, Sparks, Griskevicius, & Barclay, 2014).

Costly signalling theory (CST) and the handicap principle (Zahavi, & Zahavi, 1997) have proven to be useful frameworks for explaining the subset of behaviours related to the signalling of abilities, particularly as they relate to extravagant displays (Hawkes, O'Connell, & Jones, 2014). CST relies on the premise that only agents with a given quality (or a given level of the same) can "afford" to produce a certain signal (action), which can thus reliably communicate the possession of an underlying ability because lower-quality individuals couldn't (rationally) afford to produce it (Grafen, 1990).

As such, attending to extravagant displays can be a serviceable cue of one's ability, but it doesn't necessarily result in attributions of prosociality (Bird & Power, 2015), which are at the core of the expected benefits factor. Prosociality is different inasmuch as it can, for the most part, only be observed indirectly: there is no reliable, costly signal for one's *disposition* to help others, apart from the history of one's previous experiences with a given partner, or the information about their behaviour with others, compiled as a reputation or image score (Wedekind, & Milinski, 2000; Milinski, Semmann, Bakker, & Krambeck, 2001). Humans both actively seek this type of information (Swakman, Molleman, Ule, & Egas, 2016) and transmit it to others through gossip (Feinberg, Willer, & Schultz, 2014), to the extent that sharing reputational information can have a more stabilizing effect on cooperation and engender more trust between actors in social exchange than direct punishment (Wu, Balliet, & Van Lange, 2016). This tendency of transmitting reputational information relates to the variable of group embeddedness, by which the actor's expected benefits are influenced by how well-connected an audience member is socially, and how likely they are to transmit their belief about the actor's partner value to others. Finally, group membership has also been shown to affect prosociality in both

children (Sparks, Schinkel, & Moore, 2017) and adults (Balliet, Wu, & De Dreu, 2014), and is as such also included (indirectly) in ascertaining expected benefits of interactions.[7]

Competition for partners will, therefore, rarely be based on either prosociality (willingness to help) or ability (competence) alone. Both are universal dimensions of social perception, which influence stereotypes and emotional responses towards individuals depending on their positioning across the two axes (Fiske, Cuddy, & Glick, 2007; Cuddy, Fiske, & Glick, 2008). If ability is already apparent without incurring (however small) costs for helping, high-quality individuals might set their sights on activities which increase their fitness in different ways, like mating, or advertise their quality by choosing not to engage in signalling or impression management at all (Feltovich, Harbaugh, & To, 2002; Gambetta & Székely, 2014). On the other hand, low-quality individuals might have no other recourse than to compete by being generous (Barclay, & Reeve, 2012).

All else being equal, prosocial dispositions or 'warmth' might carry more weight than ability when choosing partners. Judgments about morality-adjacent traits (such as honesty, reliability, selflessness, or kindness) seem to precede competence judgments by being more cognitively accessible and more predictive of the valence of global impressions (Wojciszke, Bazinska, & Jaworski, 1998). This advantage should particularly be borne out in contexts where exogenous factors make the future usefulness of certain resources or abilities unpredictable, i.e. in unstable environments (markets). We predict the type of interactions one can expect with audience members to mediate the effect of ability, and consequently the expected benefits.

[7] We've already referred to group membership on several occasions, reflecting its importance in nudging prosocial behaviour. It subsumes several relevant dimensions of information: availability and probability of future interactions, expected benefits through mechanisms of expected generalized reciprocity and in-group favoritism, and so on.

For example, being a good statistician and a good hunter are both abilities individuals can differ in, which can be useful in a given situation. However, if one expects to be living in the wild, there is little added value of cooperating with a person who is a good statistician. Similarly, hunting skills are hardly pertinent to analyzing data and should not factor into decisions when one is picking their co-authors in science projects. This is to say, various environmental features can make certain abilities more likely to be useful, i.e. more likely to produce benefits more consistently than others, and should hence be weighted more in terms of audience value. Generosity has, indeed, been shown to have a larger influence than productivity on partner choice and perceived fairness (Eisenbruch & Roney, 2017), especially when coupled with environmental stability (Raihani & Barclay, 2016).

1.2.5. Exogenous factors

Apart from the direct or indirect pathways through which individual differences correlated with the probability of future interactions and their corresponding benefits affect impression management, there are exogenous factors pertaining to the environment and the market itself which can influence the importance and expression of the intermediate individual variables, and consequently also affect impression management strategies. The most obvious of these we've already hinted at, and it concerns environmental fluctuations which change the payoffs from certain abilities and shift competition to other spheres.

An example of this type of shift on the mating market is encapsulated in a detail from Ghodsee's (2018) book about sexual satisfaction in the GDR, where she writes:

> "(...) the renowned historian of sexuality Dagmar Herzog shared a conversation with several East German men in their late forties in 2006. They told her that "it was

> really annoying" that East German women had so much sexual self-confidence and economic independence. Money was useless, they complained. The few extra Eastern Marks that a doctor could make in contrast with, say, someone who worked in the theatre, did absolutely no good, they explained, in luring or retaining women the way a doctor's salary could and did in the West. 'You had to be interesting.'" (p.16).

This anecdote from the domain of sexual selection illustrates how certain qualities – even as evolutionarily relevant as resource access is to male intrasexual competition (Buss, & Schmitt, 1993) – can fall by the wayside due to changes in market composition (in this case, lack of significant variation in resource access).

1.2.5.1. Market size and permeability

Market size can determine the level of competition for partners inasmuch as it affects potential partners' outside options and thus influences their availability. The more people, the higher the probability the observer can find another (potentially better) person to cooperate with. It would thus seem that larger markets might increase the need for impression management, especially in those domains in which 'supply' might be lacking, like certain specialized skills.

On the other hand, in smaller markets – especially those with a high number of cooperators – one's risks of adverse consequences are higher, since market size determines the outside options of the actors as well as the observers. Furthermore, in tight-knit communities, information about one's defection is more likely to reach potential partners, making one's reputation more precarious. Being caught trying to dupe a partner in a small market is therefore likely more detrimental than it would be in a larger one, where the focal actor can potentially find others (to dupe or cooperate with), who are not privy to the information of their less savoury actions. It is probable that the evolved psychological mechanisms dedicated to impression

management operate on the default assumption of a small market, as this was likely the case during the period of the EEA.

Related to market size is what we call market permeability, which reflects the possibility of migration, i.e. how easy or difficult it is to switch between different markets (groups) of potential partners. While these actions might have generally required a transaction cost that was unfeasible for most individuals to pay, in certain situations - and in the case of markets which were more permeable - such a feat could've been advantageous if either the expected benefits of travelling or the cost of staying put were high enough. In situations of severe reputational damage – especially in highly competitive markets – these 'market outside options' could influence the (un)willingness to invest in reputational repair. For instance, Steele (1975) showed that an initial negative judgment about one's willingness to cooperate in communal matters increased compliance to later requests for participation in food-sharing projects. In other words, thinking community members perceive one as self-interested leads them to employ corrective strategies which make them more prone to advertise prosociality.

This depends on how damaged the reputation is: when actors perceive that others view them as incorrigible, or feel stigmatized and shamed without being given a chance to repent, they can move the other way and become even more antisocial as a consequence (Coricelli, Rusconi, & Villeval, 2014). Examples in modern society could range from switching between careers after serious blunders (e.g. doctors after malpractice suits, academics after fraud allegations) to moving to different countries to pursue the same domain in which one's reputation had previously been damaged, if one is certain their bad reputation won't follow.

The type of 'market' or culture one is socialized in likely plays a role in the perception of these outside options by influencing the number of potential partners one perceives are available, and consequently affecting strategies one will apply. Individuals in closed-off, non-permeable markets should, in our view, be more likely to invest in reputational repair than those who pay a lower cost of switching, as well as make more attempts to build reputations in other domains in which they might be competitive. Going back to Ghodsee's (2018) example from the GDR – which could, by most features, fall into the domain of less permeable markets given the difficulty of migration at the time – one simply had no other recourse than to compete by 'being interesting'.

1.2.5.2. Social rules

Social rules and norms affect both the benefits one can expect from an individual partner and the more abstract evaluations of what is considered desirable in the first place (by virtue of 'proscribing' what is seen as cooperative; but also the types of relations/contexts in which it can be manifested; see Lesorogol, 2007). Knowing what is commonly done (as opposed to merely "preached"; see Cialdini et al., 2006, for a discussion of the adverse effects of providing frequencies of undesirable behaviours to discourage them) is one way to ensure one both meets the expectations of the audience, but at the same time is not taken for a 'sucker' by following rules others casually disregard or is punished for over-contributing (Hermann, Thöni, & Gächter, 2008), especially in competitive contexts (Pleasant & Barclay, 2018). The composition of the market in terms of the ratio of cooperators and free-riders likely modulates these thresholds of what is expected (social norms), such that markets with larger proportions of cooperators set a higher individual contributions threshold for the actor to be deemed competitive, unless they can distinguish themselves in some other way.

Social information provides agents with the understanding of this state of the market, which can be used to adjust one's behaviour in whichever direction, i.e. according to the expectations of one's valuable partners. Others' donations affect what subsequent agents give, if they can observe the outcomes of previous actions. For instance, in a study by Martin and Randal (2008) which manipulated the contents of an art gallery's common (transparent) box, results showed that the composition of subsequent donations mirrored the contents of the box which the participants could see. Having information about the past frequency of contributions also nudged students at Zürich University to donate to a social fund during tuition payment in comparison to their colleagues who received no social information (Frey, & Meier, 2004), and Swiss skiers to contribute to the maintenance of a public good (Heldt, 2005).

One reason why social information influences behaviour is the aversion to disappointing (potentially) valuable partners, in the cases where their expectations are low enough or justified (Heintz, Celse, Giardini, & Max, 2015). Market norms can justify such expectations by providing common knowledge about what constitutes desirable behaviour, or what is a 'fair' share to redistribute or keep in a given context (e.g. Nettle & Saxe, 2020). In this sense, they constrain impression management behaviours to the expectations of the most valuable, or strongest audience (Marder, et al., 2016), i.e. those audience members who wield the most power to extend benefits and influence reputations on the relevant market.

1.2.5.3. Market and environmental stability

Our conceptualization of exogenous factors which influence one's benefits in a partner choice ecology also includes two closely related factors, environmental stability and market stability. By environmental stability, we mean any uncertainty or change in the *physical* environment which can influence the supply (and acquisition) of fitness-relevant resources, such

as food or shelter. In our model, we posit that environmental precariousness modulates the importance of abilities through market stability, given that it changes the payoffs from certain skills which can become obsolete when the target of the skill perishes (for example, in the case of game depletion, hunting skills become less indicative of future benefits one can accrue from cooperating with the hunter). Furthermore, market stability influences whether instances of advertising presently desirable behaviours or skills (and the resultant reputations one can acquire) are subject to change in view of value, and whether it's worth investing in them in the long-term. While a change in the valuation of skills doesn't necessarily leave one without recourse, becoming proficient in new skills is costly. For instance, when manual textile labourers started becoming redundant during the process of automation in the 19th century, the situation led to protests aimed at destroying the machines which threatened the workers' market value and reduced their competitiveness via skills – known as the Luddite movement (Autor, 2015).

How are proxies of abilities evaluated in changing environments, and how do they fare against prosocial dispositions? In a study using the Dictator game and partner choice, Raihani and Barclay (2016) showed that people prefer to choose 'poor' fair over 'rich' stingy partners for future interactions, even when there is more to be gained from cooperating with wealthier, or high-quality individuals – and that this is most markedly the case in conditions where present quality is less predictive of future quality, i.e. in unstable environments.

1.3. Proximal mechanisms of impression management

We see partner choice ecology as the key feature of an environment which could generate the selection pressures that result in proximal mechanisms sensitive to observation and consequently produce audience effects. It has already been shown to influence behaviour in economic games: contributions are higher when there is an expectation of partner choice in the

next stage of a game – beyond observability, competition for partners increases contributions even more significantly (Barclay, & Willer, 2006; Sylwester, & Roberts, 2010). Circling back to the initial examples of suspended coffee and 'free' drive-through burgers, such acts of seemingly spontaneous generosity can be explained by invoking psychological mechanisms which evolved as a means of market competition via prosociality (or 'competitive altruism'; Roberts, 1998). The fact that observation by an audience is also an inherent part of the pay-it-forward social phenomena we mention thus becomes one of the main explanatory factors which can be said to contribute to their success.

1.3.1. Error management theory and observation in partner choice ecologies

Impression management as applied to cooperative contexts should be geared toward projecting an image consistent with the standard of a desirable partner, to the best of one's capabilities. In most cases, this means being perceived as competent, concerned about others' welfare and available to help (Barclay, 2016). Put in another way, the effect of observation should induce the actor being observed to behave in a way that advertises their market value, increasing the probability of entering future beneficial interactions, or at the very least not decreasing it due to 'expensive' mistakes. What kind of mistakes can one make?

Situations which afford impression management often include a number of unknown factors. Apart from not necessarily being aware of being observed, one can also be uninformed about the relevant characteristics of the observer, such as their competence, prosociality, group membership, or group embeddedness. If one judges the latter incorrectly, for example, a socially well-connected individual could spread the gossip of one's selfishness to a large number of others, diminishing one's prospects for future mutualistic opportunities. On the other hand, if one mistakes the importance of the observer (and does this consistently), incurring costs to help

which do not lead to future benefits, one risks being taken advantage of as well as missing out on other cooperative opportunities which could add to their fitness, making them worse off in the long run. The best management of risks in impression management would've likely involved taking into consideration both of these eventualities.

A lot has been said on this account in error-management theory (Haselton & Buss, 2000; Haselton & Galperin, 2012), according to which systematic biases are to be expected in areas in which the costs of false-positives (e.g. reacting to threats when they don't exist) and false-negatives (e.g. not reacting to threats when they in fact exist) were asymmetrical during evolutionary history. Evolved psychological mechanisms in humans are predicted to err on the side of caution in domains with high fitness-related stakes attached such as impression management. Humans should avoid the choices that can disappoint desirable partners, but seize low-risk opportunities to make self-serving choices (when those are unlikely to have a negative effect on their reputation).

Accordingly, impression management mechanisms should be especially attuned to cues of observation. Given the high stakes in managing impressions, it has been tempting to claim that observation is as important as to require only minimal cues to produce large changes. An extensive body of research has dealt with the effect of such "eye" cues on cooperativeness in both experimental (Haley & Fessler, 2005; Burnham & Hare, 2007; Fehr & Schneider, 2010; Oda, Niwa, Honma, & Hiraishi, 2011; Matsugasaki, Tsukamoto, & Ohtsubo, 2015) and field environments (Bateson, Nettle, & Roberts, 2006; Ernest-Jones, Nettle & Bateson, 2011), with varying degrees of success in eliciting cooperativeness. Meta-analyses showed that while an effect of increasing cooperation seems not to be borne out of the pooled data (Northover, Pedersen, Cohen, & Andrews, 2017), there are indications – from a much smaller sample of

experiments – that eye cues can act as a powerful deterrent for antisocial behaviour (Dear, Dutton, & Fox, 2019). Earlier, theoretically-driven meta-analyses showed that the watching eyes effect is only borne out in experiments with short-term exposures to the image of eyes, whereas given time, participants are able to override the impact of the "false" cue (Sparks, & Barclay, 2013).

This ability to counteract what are automatic reactions to implicit cues is relevant for several reasons. For one, it shows that the mechanisms underlying impression management in cooperative contexts are much more complex than a simple Go-Stop rule. Secondly, it provides evidence that people are able to incorporate and use new information about the value of particular cues and inputs and adjust their responses accordingly. Several studies so far have shown that various ostensibly automatic social psychological effects – such as racial categorization – can be overridden by providing ecologically accurate inputs to the designated systems on which the cues have "piggybacked" (e.g. Pietraszewski, Cosmides, & Tooby, 2014; Williams, Sng, & Neuberg, 2016). In the absence of information on the latter however, and given the costs of false-negatives in managing reputations, we predict people should behave in a way that presupposes that the audience is potentially relevant – at least until proven otherwise like in the Sparks and Barclay study (2013).

1.3.2. Explanatory advantages of mechanisms sensitive to observation

Are the choices people make in situations where audience effects occur necessarily strategic? Impression management adapted to a partner choice ecology can be achieved through various mechanisms as previously mentioned: these range from calculated Machiavellian strategies (what we refer to as *deliberated* impression management); preferences for maintaining a good reputation in one's social circle, or preferences for positive self-esteem (impression

management aimed at the self). Why should we posit a sensitivity to observation operating through various (proximal) mechanisms instead of one mechanism which evolved to 'play fair' and consistently produce genuinely prosocial behaviour (e.g. Baumard, André, & Sperber, 2013, but see also Sperber & Baumard, 2012)? Furthermore, is a preference for fairness even necessary in addition to the mechanisms mentioned, which make up what we have called strategic vigilance (Heintz, Karabegovic, & Molnar, 2016)? Can it account for the patterns of results we see in the literature on cooperation and impression management? There are reasons to think this might not be the case.

As we point to above, an evolved preference which is not sensitive to observation or does not take into account contextual affordances and individual attributes of observers would increase the number of false-positives (behaving prosocially when it does not lead to future benefits): going for a fair distribution might overshoot, imposing unnecessary sacrifices on the actor and decreasing the fitness benefits accrued through self-serving choices. It is also unlikely that a preference for fairness would significantly reduce the number of false-negatives when contrasted with preferences for others' and self-esteem, or that the ratio of false-positives and false-negatives produced by a preference for fairness would turn in favor of fairness preferences over impression management in terms of the increases of one's inclusive fitness.

More importantly, it does not appear that people behave in this way: they use situational cues related to expectations (e.g. Dana, Cain, & Dawes, 2006; Ockenfels and Werner, 2012) and justifications (e.g. Shalvi, Gino, Barkan, & Ayal, 2015) to get larger parts of the pie when they can do so and make self-serving choices, as well as tailoring their future actions in view of their current image 'score' (e.g. moral licensing, see Blanken, van de Ven, & Zeelenberg, 2015 for a meta-analysis of the effect; see also Rotella, 2020 for a review and meta-analysis of moral

licensing with regard to observation and moral ambiguity). The bulk of the literature on 'minimal cheating' partly contradicts an account of an evolved sense of fairness and points to more specific mechanisms of both impression management and distributional preferences (for a review focused on these effects from a partner choice perspective, see Heintz, Karabegovic, & Molnar, 2016).

Apart from explaining the above mentioned lab-based experiments, our account also fares better with the cultural diversity of prosocial rules and fairness judgments. Previous cross-cultural studies have found differences in economic game contributions (e.g. Henrich et al., 2005) and punishment behaviour (e.g. Herrmann, Thöni, & Gächter, 2008). While some of these findings can be explained by different interpretations (frames) participants impose on the games themselves because of cultural norms (e.g. Tracer, 2003; Ensminger, 2000; Lesorogol, 2007), they are likely also caused by differences in what it means to be a valuable social partner in a given group.

Prosocial norms of fairness, in this view, can best be understood as a result of evolved preferences for maintaining a good reputation operating in historically specific partner choice ecologies. The cultural evolution of impression management strategies could, in some cases, lead to equilibria whose outcome is an equal distribution of costs and benefits – different markets likely included different relations between the two when it came to cooperating partners, depending on external factors such as environmental features as well as individual qualities. These exogenous factors then likely shaped the requirements for and expectations of valuable partners, and people's perceptions of the same so that – together with impression management – they resulted in social norms which emphasized specific types of distributions. Studies have in deed found that perceptions of fairness and redistributive preferences are often clustered in more

localized groups rather than vast ethno-linguistic ones (i.e. partner markets; e.g. Lamba & Mace, 2013), as well as dependent on environmental factors, such as luck (Nettle & Saxe, 2020).

Finally, a sense of fairness can emerge (in individuals) from the interaction of the above mentioned evolved mechanisms (for impression management) and enculturation. Enculturation allows individuals to have a grasp of what they can expect, given their value on the market of partners, and what others are likely to expect from them. Combined with a preference for making or maintaining a good impression, such an interaction can lead to a preference for being fair if it is aligned with the expectations of one's best potential cooperators. This process can reliably lead to acquiring a sense of fairness in all human environments. If the account from André and Baumard (2011) is interpreted as a dispositional account in the tradition of human behavioural ecology (showing that strategies are adaptive), then our own account is compatible with it: it can be considered as a complementary one that opens the black box and specifies the proximal causes of a sense of fairness.

1.4. Audience effects through a lens of adaptive impression management

In the previous sections, we presented a functional account of impression management and outlined a theoretical model of the variables which should affect an agent's fitness benefits, as well as the advantages of such a model of mind-directed preferences in partner choice ecologies. The main question we pose now refers to the types of audience effects which are suitable for the presented analyses, in other words, the phenomena which could benefit from a closer, evolutionary rendering. Some have already been given this 'evolutionary psychology treatment', like the fundamental attribution error as described by Andrews (2001) in terms of costs and benefits and error management theory. Below, we propose a (by no means exhaustive)

list of other social psychological phenomena we believe can be explained by evolved impression and reputation management mechanisms.

1.4.1. Social facilitation

Being observed has been treated as a crucial determinant (and problem) of performance as early as Triplett's (1898) studies on cyclists' outcomes, which showed improved speed in 'paced' races (either competitive or merely performed with others) as opposed to races against a clock. While observation can give us an additional performance 'boost' when completing tasks we are confident to have mastered (e.g. Bond & Titus, 1983), it can also hinder it. One of the suggested causes of the difference in social facilitation between simple (increased performance) and complex (lowered performance) tasks is evaluation apprehension – the public embarrassment actors face when making mistakes in front of others (Bond, 1982), and when the audience can see them make those mistakes (Bond, Atoum, & VanLeeuwen, 1996). This difference easily fits into our model, in that the presence of others who observe an action influences one's reputation - especially in this case of abilities.

We can further predict that evaluation apprehension would be more pronounced when i) performing in front of relevant audiences (those that can confer benefits and who one is likely to encounter again); ii) the audience doesn't have enough information about the trait X which the observed behaviour x is thought to be indicative of (e.g. intelligence); iii) the trait X is a relevant determinant of partner choice in the given context; and iv) the agent performing the action has limited outside options for future interactions. Furthermore, given recent changes in working conditions, researchers could also ask whether simply being 'seen' while doing a task is enough to elicit the effect, even when the action is not visible, and how this 'virtual' observation affects productivity; whether audience features matter in these cases as well, and so on.

1.4.2. Stereotype threat

Stereotype threat could also be explained by invoking audience expectations and the type of evaluation apprehension seen in social facilitation. When cues about the (negative or positive) expectations of an (imagined or real) audience related to one's social role are made salient, impression management mechanisms might shift behaviour closer to these expectations (whatever their valence) and cause the type of interference in information processing as posited by the theory of stereotype threat (Steele & Aronson, 1995). The relevance of audience features and observation has also been documented in this regard. In the initial experiment, for instance, the experimenters had been white men while the stereotyped group were African American students. In another study, which tested political knowledge, the race of the interviewer was found to have an effect on performance - only on the African American survey respondents, such that they made fewer errors when they perceived the interviewer to be a person of the same race than a different one (Davis & Silver, 2003). Furthermore, children from lower castes in India were shown to have lower performance in solving mazes when their castes were publicly announced, but not when they were only known by the experimenter (Hoff & Pandey, 2006).

Findings from the area of sports also point to the significance of observation: after watching videos of the 'best' free-throwers in the NBA depicted as being African American, White, or seeing a neutral video, White males shot a series of free throws, either 'observed' (via video recording) or unobserved. The unobserved players' performance was enhanced by the salience of the 'positive' stereotype, while observed players' performance in the same condition was decreased. Similar decrements were found in both observed and unobserved conditions with the 'negative' stereotype, explained by the lack of pressure to perform, i.e. of lowered (imagined) expectations which render observational effects insignificant (Krendl, Gainsburg, & Ambady,

2012). This pattern of results is strikingly similar to the above-mentioned effects of social facilitation in terms of performing complex tasks, with the added modulation of observation and performance expectations.

Stereotypes related to violating gender roles - and discomfort with consequently being misclassified as 'gay' - are also subject to modulation via audiences. Bosson, Taylor, and Prewitt-Freilino (2006) showed that audience size, familiarity (friend or a stranger) and the biological sex of the audience play a role in the perceived discomfort of violating gender roles in front of an audience. (Heterosexual) participants in this study reported more expected discomfort when imagining gender role violations in front of men as opposed to women, strangers as opposed to friends, and multiple as opposed to single observers, with the effect being larger for men. The influence of audience size reflects the predictions of our model, as does the decreased discomfort in front of friends (who have the relevant information on the trait in question, in this case one's sexuality) in contrast to strangers (who might form wrong beliefs on account of the display).

1.4.3. Self-enhancement, self-deception and overconfidence

A number of biases in the social psychology literature literature point to a tendency to inflate the relevant traits of the self as opposed to others, such as the better-than-average effect (Alicke, 1985) in which people consistently rate themselves to be higher on desirable traits than their imagined average peers, especially when these traits are controllable. Some have pointed to the potential evolutionary benefits of self-deception, as a mechanism tailored to believably deceive others about one's valuable traits while at the same time concealing this intention from the self and making it less detectable (von Hippel & Trivers, 2011). The idea of self-deception is conceptually similar to certain facets of overconfidence, namely overestimation - thinking one's

performance in a certain task is better than it is, and overplacement - thinking one's performance is better than that of others (Moore & Schatz, 2017).

Overplacement effects are likely to be more pronounced because relative performance can be difficult to ascertain for observers as well as actors themselves, which likely leaves more room to 'fudge' in this domain while not compromising one's apparent trustworthiness (this is often the form of investigating better-than-average effects). However, following the logic of our argument, actors should have access to relevant and as-true -as-possible information about their standing in a given market in order to be able to efficiently vie for partners. On the other hand, both self-deception and overestimation encounter the same problem: above and beyond appearing as trustworthy when (deceptively) reporting or advertising one's accomplishments or abilities, what is the benefit of (even implicitly) making promises one can't keep? What is the benefit of false beliefs about one's abilities, relative to the costs of being 'found out'?

We believe the weight is more likely to fall on the side of the cost, making blatant self-deception and overconfidence without access to the true information more detrimental than beneficial. This is especially true for verbal claims of (over)confidence which prove to be untrue, as well as nonverbal expressions when the same are easily falsifiable (Tenney, Meikle, Hunsaker, Moore, & Anderson, 2019). Furthermore, studies have shown that self-promotion is less likely to occur when there is a way for the audience to check the claim one makes about one's traits or performance in the future. For instance, people are less likely to exhibit the self-attribution bias – attributing their performance to ability rather than luck – when they anticipate performing a task again during the experiment (Wortman, Costanzo, & Witt, 1973). Sedikides, Herbst, Hardin and Dardis (2002) examined a similar mechanism of 'checks', i.e. accountability, in a series of experiments tailored to curb self-enhancement. They found that accountability

(anticipating having to justify an essay grading to another person) does decrease self-enhancement, and that this effect is mainly due to identifiability and the expectation of the other's evaluation.[8]

Indeed, misrepresenting one's competencies – especially those which might become important in future interactions with partners – can be seen as a form of cheating, and one that is often easily uncovered. If promoting a certain verifiable skill or competence is meant to increase the probability of future interactions with the audience because of the trait itself, then one might be better off being honest and having access to correct information. In fact, one might sometimes even benefit from under-reporting one's performance as a way to ensure justifications are present if he or she fails to live up to the partner's expectations. Self-handicapping can thus, paradoxically, be an impression management strategy (Tedeschi & Riess, 1981) – for instance, Ferrari and Diaz-Morales (2007) examined the relationship between procrastination and self-presentaton tactics and concluded it could be one way to make oneself look 'special' and the achievement even greater than it is by adding obstacles to finishing a task.

That said, we do not exclude the possibility that self-enhancement biases and true beliefs about one's competencies might be processed by different modules and thus be more or less accessible to the conscious 'I' (Kurzban, 2010). In this case, observed behaviour could be a 'joint' outcome of an impression management mechanism which it biased towards positive illusions, yet also takes its inputs from the modules in which the true beliefs are stored, thus curbing self-enhancement strategies in the domains in which they're easily falsifiable, but having

[8] Similar explanations have been proposed for social facilitation as well. Blascovich, Mendes, Hunter & Salomon (1999) developed a biopsychosocial model of challenge and threat, positing that insufficient resources to tackle a task (unfamiliarity) lead to threat response patterns, while familiarity leads to patterns corresponding to challenge situations. These responses are thought to occur during observation (as opposed to when one performs a task alone) because an audience increases goal-relevance and the value of performance to the actor (Seta & Seta, 1995).

a larger effects when the relationship between behaviour and the underlying traits is vague. Furthermore, effects connected to overplacement should also be influenced by the market and the auience one has in mind while making one's estimations. For example, self-enhancement should be more likely in cases where one expects the assertion to be evaluated by observers who are less competent in the domain than when expecting evaluations from experts.

1.4.4. The influence of audience values and beliefs

Audience beliefs and values should also direct the type of impressions one should strive towards, especially if the audiences are relevant to one's future well-being. Schneider (1981) points to the importance of meta-cognition when talking about the effectiveness of impression management – stressing the knowledge about the average target (i.e. audience member) as crucial in determining which from a large repertoire of possible behaviours and strategies one should pick in order to achieve the desired impression.

One need not always aim for the best impressions: under-performing can also be a strategy in itself. For instance, Zanna and Pack (1975) found that women underperformed on a purported measure of intelligence if their expected audience was a highly desirable male student with gender-stereotypical attitudes (as well as reporting more conventional attitudes), but not when the potential partner was undesirable or did not have such attitudes. Importantly, their design included hints about the student's availability and wish to meet others, presumably also increasing the perception of the probability of future interactions and availability for partnerships.

Berger and Rodkin (2012) investigated prosocial and aggressive behaviours of early adolescents who changed peer groups, finding social mobility to be an important factor.

Matching the predictions we would make from the theoretical model outlined in the previous sections, the social norms regarding prosociality and aggression of the 'new' attracting-groups (the ones participants were joining) had a larger influence on these behaviour than the norms of the departing-groups (those the participants were leaving). Malkis, Kalle and Tedeschi (1982) showed that students tend to only feign attitude change (so as not to appear inconsistent or dishonest) when the experimenter is a researcher from their own university (one they might encounter in the future), but not when they are from a government department and thus unlikely to be relevant in future interactions.

Alexander and Weil (1968) showed that when winners and losers of a Prisoner's Dilemma game were differently judged by the audience (either positively or negatively), the rate of cooperative choices increased significantly if the audience's judgment of the 'loser' included positive traits such as friendliness or generosity. While the majority of studies find an increase in prosocial behaviour when it is made public, Dufwenberg and Muren's (2006) data thus went in the opposite direction: they found participants behaved less prosocially in public than in private. However unexpected the result had been, their attempt at providing an explanation was intriguing on several accounts: that the counterintuitive outcome was due to the sample being Economics students, who are taught to value the ideal of selfish, rational maximizers and want to portray themselves to their peers as such. It underscores the assertion that reputations need not be "good" to be valuable, and that the values of the audience can have a significant impact on prosocial behaviour.

Going to the extreme of what types of impression might be adaptive, Diego Gambetta (2009), in his ethnography of signalling in the Italian mafia, notes that aggression and even murder can be a form of costly signalling (and a way to make a favorable impression):

> "First, being suspected of murder, a stigma in polite society, can be a bonus in the underworld. Whether it is good or bad depends on the audience one aims to impress. Next, being suspected of a serious criminal offense has a peculiar by-product: it provides hard-to-fake evidence that one is "bad," and it spreads the knowledge of this trait, which is arduous to advertise both credibly and widely otherwise." (p. 59).

In the context of the mafia, these types of cues can be viewed as a form of adaptive impression management. By managing one's impressions in the direction of a sub-community (the mafia), and especially in a milieu in which strategic vigilance plays a key role in partner choice because of the constant risk of being caught by the police or cheated by associates, one incurs a large cost in terms of outside options (the members of 'polite society' from the quote above), thus showing commitment to potential partners from the same group.

Finally, humans often comply with the views of large groups, even if we suspect their decisions are wrong (Asch, 1951, 1956). This is especially true when reporting is done publicly in front of the same group of observers, rather than privately – a difference that has been conceptualized as 'compliance' (to a false belief) rather than its 'acceptance' (Sowden et al., 2018). Audience size, combined with uniform behaviour, can also provide more social information on the expected conduct in a given situation. For example, a meta-analysis of conformity as it relates to group/majority size, as well as public and private behaviour, has shown a small, but positive significant relationship for public responses between conformity and majority size in Asch's line task (Bond, 2005).

1.4.5. Adaptive impressions

What kind of impressions do people usually aim to achieve? While it is tempting to claim they prefer the best, or at least positively valenced, impressions, this doesn't have to always be the case. Jones and Pittman (1980) considered five distinct types of impression management strategies, including ingratiation, intimidation, self-promotion, exemplification and supplication into their taxonomy. Ingratiation would be what most people think about when they think about impression management – the various tactics employed to promote *liking*. On the other hand, intimidation is used in order to threaten the audience into complying to a certain behaviour which fits one's goals. Self-promotors aren't as concerned with liking as they are with advertising their competence and skills, whereas exemplifiers seek out approval for their moral virtues. Finally, supplication as an alternative to ingratiation is used by deference to one's superiors and highlighting one's weakness in order to gain favor (perhaps by counting on the other's pity or sympathy, or conveying one will be satisfied with a smaller ratio of the jointly accrued benefits). Though by no means exhaustive, the resultant impressions from the various strategies make it clear they need not always be positively valenced – the process is different than merely pleasing the audience (Baumeister, & Tice, 1986).

People might not always aim for acquisitive self-presentations (gaining social approval or leaving a certain impression), either: sometimes one might be better served by trying to *avoid disapproval* or employing a protective self-presentation style (Arkin, 1981). Actively trying to ensure the best impressions is not always desirable as it can make one a target for do-gooder derogation and even punishment – especially in biological markets driven by competition (Pleasant & Barclay, 2018), or those with low tolerance for deviating from social norms (Kawamura & Kusumi, 2020). In cultural groups where social norms emphasize humility or

uniformity, as exemplified by the Danish 'Law of Jante' (Cappelen & Dahlberg, 2018) or the 'tall poppy' syndrome (e.g. Peeters, 2004) – one might be better off employing a conservative orientation in self-presentation (characterized by modesty, neutrality, conformity and compliance; Arkin, 1981) instead of being metaphorically 'cut off' first in the garden. We expand on the possible benefits of 'opting out' of prosocial impression management in the next section.

Finally, one need not employ directly self-enhancing strategies at all. Basking in reflected glory (Cialdini et al., 1976) and cutting off reflected failure (e.g. Boen, Vanbeselaere, & Feys, 2002) are both strategies which operate via one's relationship to a group (or a social identity). In the first case, people are more likely to claim association with popular and positive groups, whereas in the second, they aim to distance themselves from an identity or group when it is negatively valenced (e.g. in the case of losses by one's sports team). Finally, 'blasting' (Richardson & Cialdini, 1981) – derogating those identities, people or groups with which one is negatively associated or in direct competition – and 'boosting' (Finch & Cialdini, 1989) – increasing the positivity of a negatively valenced connection by extolling some of its traits – also belong with the above as indirect strategies used to manage impressions.

The benefits of these can sometimes be three-fold. For one, elevating a group one is associated to is also expected to elevate one's own standing with the audience, in an inverse of the 'guilty by association' idiom. By making one's relationship to a successful or popular group known, one expects to score 'impression points' for oneself. Secondly, by advertising one's association to a certain group, one can also gain those points from the group itself, by affirming one's loyalty and dedication to the group. When it comes to blasting, one again indirectly enhances one's own impression by decreasing the desirability of the other, rival association,

which can come in handy on a biological market where reputations are relative and one is competing with a large number of others for access to both valuable partners and high standing. Thirdly, cutting oneself off from things or groups which suddenly become low-value can be seen as a way of avoiding the above mentioned guilt by association (by removing the cues of the association itself from the audience's mind), while boosting can be seen as a reputational repair strategy when this association is already known and one has to make the best of what one has to work with.

All of the above is to say that there are plenty of instances in which one might reasonably be expected to employ tactics other than impression management as it is understood in its narrow sense, and largely because what makes an 'ideal' partner varies across contexts and situations, or, more broadly, markets. Which strategy one is likely to use should depend largely on the present context and the kinds of behaviours it affords for managing impressions, but also on the type of expected interactions with the audience. Given the myriad situational cues which need to be taken into account, as well as individual differences, it is no wonder that impression management often fails to produce the desired effect (for an account of people as impression 'mis-managers', see Steinmetz, Sezer, & Sedikides, 2017). In the following section, we argue why this might especially be true in the prosocial domain,

1.5. The Catch-22 of prosocial impression management

So far, we've mostly dealt with the factors and intermediate variables which influence fitness through impression management, focusing on the side of actors' probability of strategic investment in reputation. However, as partner choice mechanisms likely evolved in an arms' race with mechanisms dedicated to parsing who the best partners actually are, audience scepticism about the informational value of the observed action should also increase with the perceived

motivation to manage impressions – in particular when it comes to prosociality which, as previously noted, can mostly be inferred only out of a given context and the embedded action, rather than directly assessed.

To be valuable, an evaluation of a potential partner needs to be generalizable and informative of the decisions they will make across situations, regardless of whether strategic incentives are present or not. The reputational benefits embedded in the context of an observed prosocial action are therefore to be taken into account when making prosocial attributions. In other words, "choosers" should be strategically vigilant (Heintz, Karabegovic, & Molnar, 2016); the audience should be "sceptical" (Bird & Power, 2015). The same situational variables which lead to more signalling will in this view lead to more scepticism and vice versa: this arms' race should lead people to anticipate others' scepticism when making the decision whether to compete via prosocial choice, especially when it comes at a high cost. Since the exact combinations of variables which influence outcomes in different situations are unlikely to be repeated in identical ways, taking only the final product of an action is a poor basis on which to make decisions about whom to trust. A better strategy involves computing the underlying intentions and attributing dispositions (or traits) which reflect intentions and outcomes separately. This ensures better predictions of both future intentions and subsequent behaviour, and lowers one's chances of being duped.

The underlying assumption of considering a mechanism dedicated to attributing prosocial dispositions is that there are, in fact, stable prosocial tendencies to attribute. In other words, that people who behave prosocially in one situation will do so in others; that there are inherent differences across the population in the intrinsic motivation to repeatedly provide benefits to others at a cost to oneself, as opposed to either opportunistic helping or selfishness. This

tendency can be a personality trait, such as introduced by the HEXACO model's honesty-humility dimension (Lee & Ashton, 2004), a prosocial value orientation (Van Lange, Bekkers, Schuyt, & Vugt, 2007), or a social preference (Charness, & Rabin, 2002) – the specific definition of what it is that drives the behaviour isn't crucial. What is important is the existence of variation in the likelihood of prosocial behaviour between individuals, and that certain combinations of environmental factors have more influence on some individuals as opposed to others. This is to say, if one wants to choose the best partner available, the key adaptive task that needs solving is distinguishing not only who is likely to cooperate and not defect, but who will do so across a wider variety of situations, and despite possible temporary appearances that one might not be able to reciprocate in the near future (for example, due to sickness or injury).

There are studies which show that people act consistently over time and across situations, and that personality can predict behavioural consistency even after situational similarity is controlled for (Sherman, Nave, & Funder, 2010). Active cooperation has been linked with the abovementioned Honesty-Humility dimension of the HEXACO model (Hilbig & Zettler, 2009), with studies also showing that punishment predominantly affects only the contributions of individuals low on the H-H scale, such that they condition their contributions to the public good on the existence of external factors, unlike those with high scores (Hilbig, Zettler & Heydasch, 2011). Peysakhovich, Nowak and Rand (2014) tracked the behaviour of online participants in different economic games, repeated after a period of 3 months, and found contributions to be stable across games which rely on cooperation (as opposed to punishment), and across different time points. Yamagishi et al. (2013) also showed consistency in decisions between five economic games and across a large time interval (3.5 years), as well as the the importance of how the game situations were perceived (as collaborative or otherwise) for subsequent

behaviour. These results indicate that attributing dispositions relating to cooperativeness can be a good internal mechanism for predicting future behaviour, especially if one also takes the context into account (see also, Thielmann, Spadaro, & Balliet, 2020, for a focused meta-analysis of personality traits, prosociality and situational affordances).

Since there seem to be individual differences in prosocial dispositions, and since they are predictive of a wider range of future behaviour – observers should take special note of the possible strategic motivations behind observed prosocial actions. This means again taking into consideration the same factors such as audience size, presence of high-status individuals or those otherwise relevant to the actor's goals, as well as what the actor knows about the situation (e.g. whether they are aware of being observed) or how they go about the prosocial display (for example, in an exaggerated, ostentatious way meant to emphasize the action, or a less conspicuous way). Regarding the latter, Bird and Power (2015) found that pecuniary distancing – sharing in a way that "disengages" the actor providing the goods from their subsequent distribution – is linked to prosocial generosity, and that incurring larger costs to provide benefits for the community has more impact on the formation of lasting, cooperative partnerships than status-enhancing displays which invite scepticism about prosocial actions.[9]

Recent signalling models have provided additional insights about the cues that are attended to when evaluating cooperativeness. The game-theoretic "cooperate without looking" paradigm (Hoffman, Yoeli, & Nowak, 2015) was modelled to show that cooperating without calculating the cost to oneself can be beneficial to the actor if "blind" cooperators are trusted

[9] This seems to be an effect specific to advertising prosociality, as studies on bragging have shown that emphasizing one's competence-related traits influences liking of the target agent, but not the evaluation of the trait being bragged about, as opposed to prosociality where it has an effect on the target trait as well (Berman, Levine, Barasch, & Small, 2015).

more by others. The speed with which one makes the decision to act prosocially is another related feature which can cue this uncalculatedness, the underlying reasoning being that people will make their "default" choices faster, possibly driven by a set of internal preferences. Jordan et al. (2016) empirically tested both hypotheses and showed that reputation effects do drive uncalculating cooperation, that is, that people make cooperative decisions both more rapidly and are less likely to "look" at the cost of helping when being observed by a future partner, which is rewarded by the partners transferring a larger part of their initial endowment in a subsequent Trust game. Gambetta and Przepiorka (2014) also showed that generous choices made by uninformed players are more effective in engendering trust than generous choices which are potentially strategic.

Someone whose choice is being observed is thus better off hiding their goal of being judged positively, which sets an important challenge: how to convey one's prosociality, if doing so intentionally will be interpreted as self-serving? The above-mentioned studies imply that actors are sensitive to ways in which they can make prosocial decisions seem more genuine (i.e. by not looking at the cost of the action or by making it quickly). However, the question of how they attempt to overcome seeming self-serving when their strategic motivations are made apparent is less clear.

The success of "cost-blindness" is contingent upon whether or not the ulterior motive is retrievable to the observer. Since a lot of social interactions include a conflict of interest, overt advertising of prosociality (publicly and intentionally making one's past prosocial choices known) might understandably be met with a dose of scepticism, wherein the retrieved intention to convince others of one's value as a cooperator overshadows the intention to produce a benefit for the target. There are a number of additional effects in the literature on evaluating altruism

which show a similar pattern of distrust, even at the slightest changes in context. Having a personal connection to a certain cause can thus "cheapen" one's altruistic behaviour and affect observer's ratings of agents' intrinsic motivation, even when they have nothing to gain by being involved in the cause (Lin-Healy, & Small, 2012). Accruing benefits from a charitable action also leads to attributions of self-interested motives, even when said benefits are randomly obtained and could not have been the strategic goal underlying the action on the actor's part (Lin-Healy, & Small, 2013). Producing beneficial outcomes for others as a consequence of pursuing a self-interested goal is similarly viewed as worse than not providing any benefits at all (Newman, & Cain, 2014). Fein and Hilton (1994) note that when suspicion about one's possible ulterior motives is raised, observers seem to devote an inordinate amount of cognitive effort into parsing the context of the situation and coming up with alternative explanations for the observed action (in their case, advocating for a certain position). Given the fitness consequences of wasting valuable resources such as time and effort into pairing up with less cooperative partners, it isn't surprising that humans should be wary when assessing ambiguous situations, especially in this domain.

The question remains open as to the extent to which "cheapened" altruism is truly cheapened, and under which conditions. It is possible that prosocial actions are simply "discounted" or taken with a grain of salt, dependent also on individual or audience levels of scepticism or self-interest. In this case, the action would still influence the impression of the target agent in a positive way, only the net benefit from it would be lower than had it been embedded in a less suspicious context. Another possibility is that prosocial actions heavily implied to be strategic are not taken into consideration at all when attributing prosocial dispositions, but discarded for lack of their informational value for the question at hand (i.e.

since it is impossible to gauge how the agent would behave in a different context, it is neither used to update their prosocial reputation positively, nor negatively). Finally, the possibility also exists that strategic investments in prosocial reputation, if judged as such, would have a self-defeating effect and actually *lower* the observer's evaluations of one's prosociality, similar to the tainting influence of benefits on judgments of altruism.

1.6. Conclusion

The primary goal of this chapter was to provide a basis for an evolutionary psychology-inspired program for audience effects by outlining the relevant factors and cues which should serve as inputs to an evolved impression management mechanism. While some of the classical studies employed functional terms when describing the prerequisites and goals of self-presentation, they rarely ventured outside of the scope of psychology as such in their search for explanations. Our second goal was to provide a (limited) overview of different social psychological phenomena which might fall under this same umbrella, vis-à-vis their ultimate function. We believe that the effects mentioned in this chapter fall neatly into the predictions of a BMT-based model of audience effects, and that they could benefit from future investigations in view of these proposed evolutionary origins. Finally, the outlined model will serve to inform the rationale behind the experiments we discuss in the rest of the thesis, which rely heavily on the ideas of adaptiveness in impression (mis)management, audience features and beliefs, and intentions in the prosocial domain.

Chapter 2: Credible evidence in prosocial displays

2.1. Introduction

Impression management is of special importance in interactions which include high stakes for the future benefits of the 'impression manager', such as job interviews with potential employers (Bourdage, Roulin, & Levashina, 2017), which can be viewed as a particular instance of partner choice. While the desired type of impression can change depending on the circumstances (e.g. if one prefers to present oneself as competent or prosocial), the motivation to influence the beliefs of others in a way that is beneficial for oneself rarely does. Reaching one's goal through impression management can thus come about through different strategies, as outlined in the previous chapter. These can involve either deliberate (conscious) strategies or less deliberate processes like self-enhancement biases (Sedikides, & Gregg, 2008), which include an array of different effects. For instance, people can evaluate themselves as possessing more of a desirable quality than an 'average' peer (Alicke & Govorun, 2005) or be more likely to ascribe successes to internal causes, and failures to situational factors beyond their control (Andrews, 2001, Libby & Rennekamp, 2012).

Impression management does not go rampant, however: Actors anticipate that the audience will be vigilant towards their intentions (Heintz, Karabegovic, & Molnar, 2016), especially in the domain of prosociality, where a vigilant audience can interpret a prosocial choice as motivated by self-interest rather than by genuine prosocial preferences. Monetary incentives thus often crowd out public prosocial behaviour because they 'dilute' the relevant signal (Ariely, Bracha, & Meier, 2009; Barasch, Berman, & Small, 2016) – in which case, the desired impression will not be achieved.

2.1.1. The benefits of retrieving intentions

There are important benefits in assessing the true incentives behind observed choice for the observer, i.e. in exercising *strategic vigilance*. The goal of such cognitive processes is forming an impression of others so as to accurately predict whether they will make prosocial choices in the future, especially towards oneself. Exercising strategic vigilance should lead to up-to-date beliefs about an agent's reliable prosocial dispositions and answer the following question: Will this person act in a prosocial manner when the incentives to manage impressions are absent?

In other words, strategically vigilant observers will attempt to tease apart whether the observed choice is the result of stable dispositions, as opposed to being sensitive to *volatile* aspects of the situation. If the situational features which make impression management valuable are remain stable, observers should be more willing to interact with the same person again, given they are likely to behave in a predictable way. However, if such features are volatile, one is better off not interacting with a person who was driven towards cooperation by reputational concerns instead of a stable prosocial disposition.

One-off outcomes can vary for many reasons, unconnected to the actor's intentions, whether it be through an accident or a side-effect of another decision. As the variables which influence outcomes in different situations are unlikely to be repeated, taking only the final product of an action is a poor basis on which to make decisions about whom to trust in cooperative situations. A better strategy involves computing the underlying intentions and attributing traits reflecting both intentions and outcomes (such as prosociality and its different facets as well as competence).

For example, if a colleague offers help with a statistical analysis of a given set of data, and they somehow receives a corrupt file, the fact that the final outcome is not beneficial to us

does not reflect on their intention to help and our evaluation of them as a good cooperative partner nor their competence. On the other hand, if they make an accidental error in the analysis, the wrong outcome may reflect poorly on our perception of their competence, but not on the perception of their kindness or willingness to help. Finally, if they are working on the same data as we are and as a consequence produce the results that we need; our perception of their willingness to help would most likely remain unchanged, since the contribution to our work in this case was unintended. Delton and Robertson (2012) examined evaluations of similar situations in their study of social foraging, where costs suffered by the actors to accrue benefits for the group were given as diagnostic of the willingness to help. They found that people categorize those who provide more benefits for the group (and consequently oneself) by intentionally incurring a cost separately from those who do so incidentally. Participants also evaluate the former as more altruistic and report they would be more desirable partners in similar situations.

We enrich the context of the Delton and Robertson (2012) study by adding a distinction related to the strategic incentives present in the context in which the behaviour takes place. Think of the aforementioned situation in which we need help with statistical analyses, only this time we make it known in front of the principal investigator of the project. A colleague offers to help, and we assume all else goes smoothly. In this scenario, the conclusion that the colleague is prosocial and would help us with future statistical endeavors is not as straightforward, because there is also the possibility they might simply have offered help so as to 'show off' with the PI.

There are various factors which influence whether an observed prosocial choice will lead to an attribution of prosociality, observation being one of the key variables we're interested in. De Freitas, DeScioli, Thomas, and Pinker (2019) investigated impressions of charitability at

different levels of anonymity and common knowledge, showing that both variables play into observers' impressions of the actors. Specifically, those who make charitable contributions while advertising their identity are generally seen as less charitable than those who choose not to make themselves known as the do-gooders. De Freitas et al. (2019) found that even subtle variations in the way knowledge about each others' identity was revealed influenced subsequent evaluations, which the authors interpreted as a sign of very sophisticated psychological mechanisms underlying attributions in this particular domain. An adaptive response to observers' fine-grained appraisals of prosociality involves anticipating their lack of candor. This is especially pertinent when impression management is costly, as is the case of prosocial choice.

Another example is the conspicuousness of the choice or, at the other extreme, pecuniary distancing – people are only sceptical toward the former, aggrandizing displays (Bird & Power, 2015). As mentioned in Chapter 1, the speed with which the choice is made and the information the actor has about the situation should also be attended to: making a prosocial choice faster and not calculating its costs can be more advantageous than deliberated decisions (Hoffman, Yoeli, & Nowak, 2015; Critcher, Inbar, & Pizarro, 2013) and observation leads to using this strategy in economic experiments (Jordan et al., 2016). This is further evidence that agents manage the reputation effects of their actions by eliminating evidence of selfish motives. Are there other aspects of the situation, except the abovementioned, that observers take into account when deciding whether a prosocial *choice* was motivated by a prosocial *disposition*?

2.1.2. Rationale and hypotheses

In the following studies, we assess how observers evaluate actors making a prosocial choice in view of their interest in conveying a prosocial impression. We also look at whether

actors anticipate the consequences of embedded strategic incentives, modulating their choices accordingly. Our main hypotheses can be summarized as follows:

H1: Audience discounting. The audience assesses the likelihood of a prosocial choice being motivated by genuine prosociality or material benefits expected from making the desired impression. In the latter case, they discount their attributions of prosociality.

H2: Impression management modulation. Actors are aware of the manner in which vigilant audiences discount attributions of prosociality, and they modulate their behaviour accordingly.

H3: *Modulated audience discounting.* The audience discounting effect in evaluations of prosociality is influenced by the perceived benefits an actor can accrue from future interactions with their observers, i.e. the relevance of the audience.

We test these hypotheses by means of manipulating some of the contextual factors which affect motivations to create an impression of prosociality, and consequently observers' attributions of prosociality. We examine impression management in the prosocial domain with two methods: (1) a lab experiment with monetary incentives, and (2) vignettes with 'slices of life' stories to elicit judgments of prosocial choices, combining these two methods to increase the robustness of our main findings. The scenarios evoke familiar, every-day contexts and provide ecological validity to the effects we document, while the experimental game uses monetary incentives so as to provide material stakes in choosing an impression management strategy. In both studies, we vary the circumstances and the incentives for conveying a prosocial impression.

In Study 1, we consider prosocial choices that can stem from either the motivation to attract partners for future beneficial interactions or a prosocial disposition. Participants first make

their decisions without the knowledge of strategic incentives, and are then allowed to revise their choice (after they are made aware of the benefits of making a good impression). We implement a condition where this change is common knowledge—i.e. known to the observers—and a condition where it remains private—i.e. unknown to the observers. We document both the choices of the observed participants, who can benefit from making a good impression, and observers, whose benefits depend on selecting a partner with 'genuine' prosocial dispositions.

In Study 2, we further examine the evaluations of actors who changed (or didn't change) their initial decisions after receiving information that their choices will be observed. We also include more complex contextual factors, such as the presumed relevance of the audience for the actor and the cost of the action as independent variables, to explore the sensitivity of the underlying cognitive mechanisms to situational differences: we test participants' sensitivity to the modulation of the cost and the expected benefits of a prosocial choice that stem from impressing the audience. Furthermore, we test the modulated audience discounting hypothesis (H3) by introducing differences in audience relevance to the protagonist. We predict that the relationship of the observers to the actor and their instrumental value can either strengthen or lessen the discounting effect as not all audiences are equally worth impressing.

2.2. Study 1: Prosocial impression management under suspicion

The material benefits derived from impressing others are hard to measure in the field, but can be controlled in lab-based economic games. Barclay and Willer (2006) and Sylwester and Roberts (2010) ran repeated games with opportunities to choose partners. In such set-ups, prosocial choice can provide immediate material benefits because it increases one's probability of being chosen for future interactions.

In this study, we made use of the same methodology to test predictions about strategic vigilance. We introduced a manipulation that varied how much information the audience had at their disposal across conditions and assessed how it influenced participants' willingness to increase the contributions that are to be observed by said audience. Our setup consisted of an economic game with a first stage in which one portion of participants played a 2-person Public goods game (PGG), and another portion observed the choices made in the game. In a later stage, the observers selected a partner for a new round of the same game and they could base their selection on the observed choices from the first stage.

However, we varied the information available to the observers, so that the incentives to engage in impression management were either made apparent or not. More precisely, all actors in the first stage were given the opportunity to change their contributions after learning they were going to be observed by potential future partners. Importantly, the making the change – which provides evidence for a motivation to impress – was either disclosed to the observers in the 'common knowledge' condition, or kept secret in the 'private knowledge' condition (for a similar experimental manipulation in the domain of emotions, see Andrade & Ho, 2009).

According to our impression management modulation hypothesis, actors should be aware of how conspicuous displays of prosociality come across to an audience when the strategic dimension of the choice is made salient. Because observers are privy only to the information that a change was made and the changed amount in the common knowledge condition, changing one's contribution leaves them to speculate on the initial decision, which would be more informative about genuine prosociality (if no change is made, the initial transfer remains known to observer). They thus need to apply their intuitions about impression management to the final decision. Similarly, actors need to keep in mind the extra 'cost' of suspicion which comes with

changing their contribution when the change is known, and adjust their decisions accordingly. We made the following predictions for the choices of actors and observers in Study 1:

P1. Increasing one's contribution will be more frequent when the observer isn't aware of the strategic incentive underlying the change for the actor group (derived from H2).

P2. IM by increasing one's contribution will be perceived as beneficial only to those with very low initial contributions in the common knowledge condition, making the differences between initial and changed contributions larger in the common knowledge condition (derived from H2).

P3. Observers will prefer to interact with actors who did not change their contributions as opposed to those who did, even when the changed contributions are slightly higher (derived from H1).

2.2.1. Method

2.2.1.1. Participants

Three sessions were held for each condition (9-18 participants), with the final number of participants amounting to 87 (40 female, 47 male, *M* age = 26.90): 42 in the first (28 actors and 14 observers), and 45 in the second condition (30 actors and 15 observers). A power analysis based on a medium-to-large difference (d=.70) in means of change between the two actor groups, with alpha set at .05 and statistical power set at .80, rendered a sample size of 34 per group (Faul, Erdfelder, Lang, & Buchner, 2007), however, we had to stop at the sample size given due to external constraints in participant recruitment. Sample size was determined before any analyses were conducted. We additionally gathered data from 35 observers (17 female, 3 other-identifying, 15 male, *M* age = 25.62) as a follow-up to test H1. Out of the 35 participants we

recruited, we excluded data for 7 of them who had answered all the comprehension checks regarding the understanding of the game incorrectly.

All participants were recruited through the SONA recruitment system at the Central European University, with the condition of proficiency in the English language. The experiment was approved by the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

2.2.1.2. Procedure

The experiment was programmed and carried out using the Z-Tree software (Fischbacher, 2007). Participants interacted with each other using desktop computers at the University 'Green' and 'Blue' computer labs, which were partitioned off to provide anonymity. Upon arrival, they were randomly divided into groups by pulling cards with "green" and "blue" monsters from a box (analogous to the names of the two labs). Participants who pulled the blue card were given the roles of ''actors'' and stayed in the larger computer lab: this was the group which would play the game first and whose contributions would later be observed. The green group were given the role of "observers" and moved to another room with another experimenter: this was the group that would observe the interactions and choose their partners accordingly. In the beginning of the experiment, participants were not informed about the task of the other group and no mention of possible future interactions was made.

Stage 1: Two-person Public Goods Game

In the first step of the experiment, the 2-person Public goods game was explained to both groups, and a payoff matrix was provided on a sheet of paper, containing the pre-calculated earnings for both actors and each possible pair of contributions. Every actor in the Blue lab got

an initial endowment of 10 money units (MU), which they could invest into the common box shared with a partner. The investments of both actors in the pair were multiplied by a factor of 1.5, and redistributed equally at the end of the round. Actors were told they would play three rounds of the game with the same partner, only one of which would count toward their payoff. Observers were additionally informed about their task of choosing a partner to play the game with, and the information they would receive to base their decision on. All participants were then asked to complete a 3-task comprehension check, and given detailed feedback of how the correct answers are calculated after each task to ensure understanding. Stage 1 of the experiment was the same across conditions: the Blue group played three rounds of the 2-person PGG with their designated partner and was given information about their earnings after each round.

Stage 2: Information about partner choice

Stage 2 began after actors in the Blue group received feedback about their earnings from the third round. At this point, the experimenter explained that their contributions from the third round (R3) would be observed by the participants in the Green group, who would be able to choose one actor out of a pair to play an additional (paid) round with them. It was also revealed that the earnings from R3 would be used to calculate their payoffs at the end of the experiment. The option of changing the R3 contribution was introduced, i.e. participants could "re-play" R3 after being made aware that the contribution would be observed.

The two conditions differed in what the observers would know about the contributions they saw. In the first, private knowledge condition (C1), actors could alter their contributions without the observers having the information of the possibility of changing – they were simply presented with the information of the final contributions from the revised R3, a fact of which

actors were aware. In the common knowledge condition (C2), observers were given the additional information about whether the actor had chosen to alter their contribution after being notified there would be a selection process for another game (though they would not be aware what the previous R3 contribution had been). Again, actors were made aware of the information observers would be privy to.

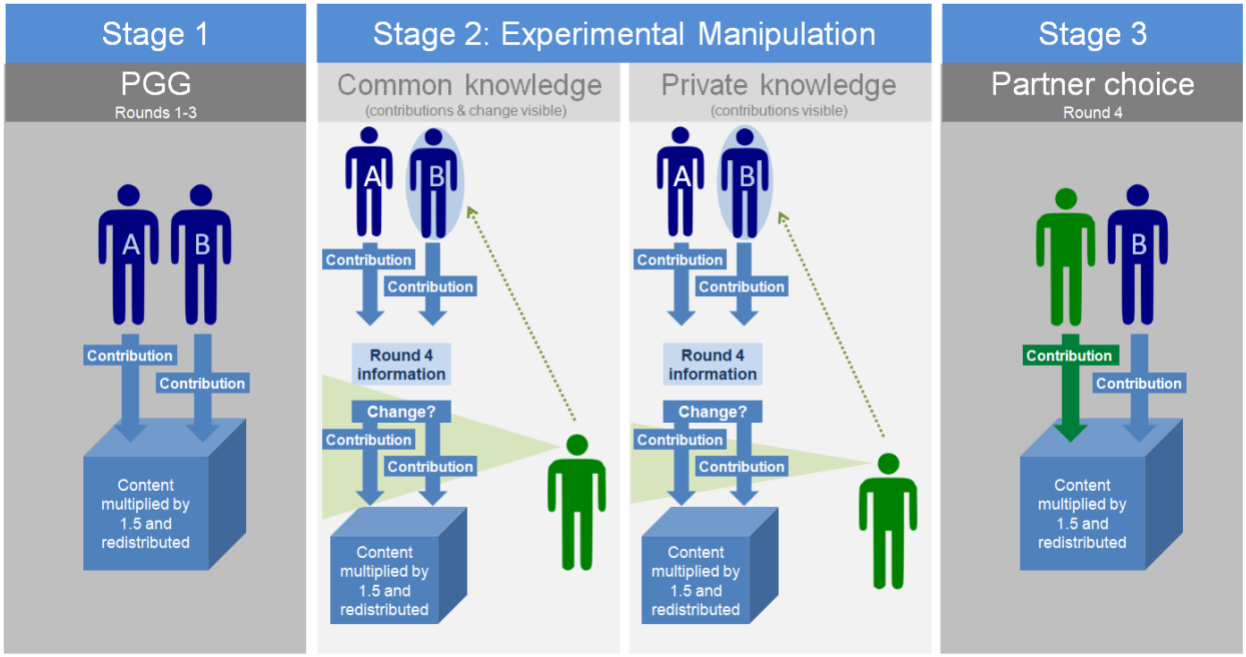


Figure 2.1. Experimental setup and manipulation, Study 1. Stage 1: Actors in the Blue group play three rounds of the two-person PGG. Stage 2: Actors are told about the subsequent partner choice round and given the option to change their contribution from Round 3; depending on the condition this change is either visible to the observer or not. Stage 3: Observer picks one out of a pair of participants and plays another round of the PGG with the chosen partner.

Stage 3: Partner choice

In stage 3, observers were presented with the information about the Blue group's task and the contributions of a pair of actors, with the information about the change being hidden (C1) or displayed (C2), according to condition. They then chose the actor they wanted to interact with in

a 2-person PGG, which consisted of one paid round, and played the game with their chosen partners.

We recruited additional participants to the observer group, as it was only a third of our sample. These additional observers were given information about the game and the context in which actors made the decisions about contributions, analogous to the information given to observers from the common knowledge condition (C2). They were shown pairs of actors with changed and unchanged contributions ("revised R3"), and had to pick the actor they wanted to interact with. Specifically, they were told that a round had been chosen to provide their payoff from two blocks of the experiment (each constituting 38 partner-choice decisions). The rounds were randomly pre-selected among the pairs of contributions for which there was information about actors' contribution in the additional game (R4). Participants received the chosen actors' contributions from the previous R4 games.

The partner-choice pairs included encompassed all possible combinations of changed contributions being higher than the not-changed contributions (55 items), in addition to a set where both contributions were equal (11 items), and a set where the unchanged contribution was higher by 1 MU (10 items). Thus, participants were presented with 76 choices in total and had to select the partner whose subsequent contribution they would prefer to receive.

Exploratory measures

At the end of the session, participants completed the SDS-17 scale (Stöber, 2001), a measure of social desirability (and impression management), as part of an exploratory analysis into whether the tendency to give socially desirable answers would play a role in the decisions to change one's contributions because of possible experimental demand; or influence decisions

during the three rounds of the Public goods game. The measure is composed of 16 self-report questions, answered as true (coded 0) or false (coded 1), so that higher scores reflect more pronounced social desirability concerns.

2.2.2. Results

2.2.2.1. Proportion of change across conditions

Figure 2.2.A shows the distribution of strategies for partner competition used in each condition, where the crucial switch can be seen in the high frequency of small changes and the low frequency of big changes in the private knowledge condition, and the opposite trend in the common knowledge condition. Overall, 60.7 % of participants changed their contributions after R3 in the private knowledge condition, while 43.3 % did so in the common knowledge condition: this difference was not significant (N=58, χ^2=1.752, df=1, p=.186, Cramer's V=.174). A sensitivity power analysis of the study for detecting differences in proportions using G*Power (Faul et al., 2007), with alpha=.05 and statistical power set at 80%, showed the critical to be χ^2=3.841

We ran an additional analysis on the frequency of the categories of change across four intervals which mirror different possible strategies: negative change (lowering one's contribution), no change, small change (up to and including a 3 MU difference) and big change (higher than 3 MU). The difference in the frequency of participants using each strategy across the two conditions proved to be significant (N=58, χ^2=12.305, df=3; p=.004; Cramer's V=.461).

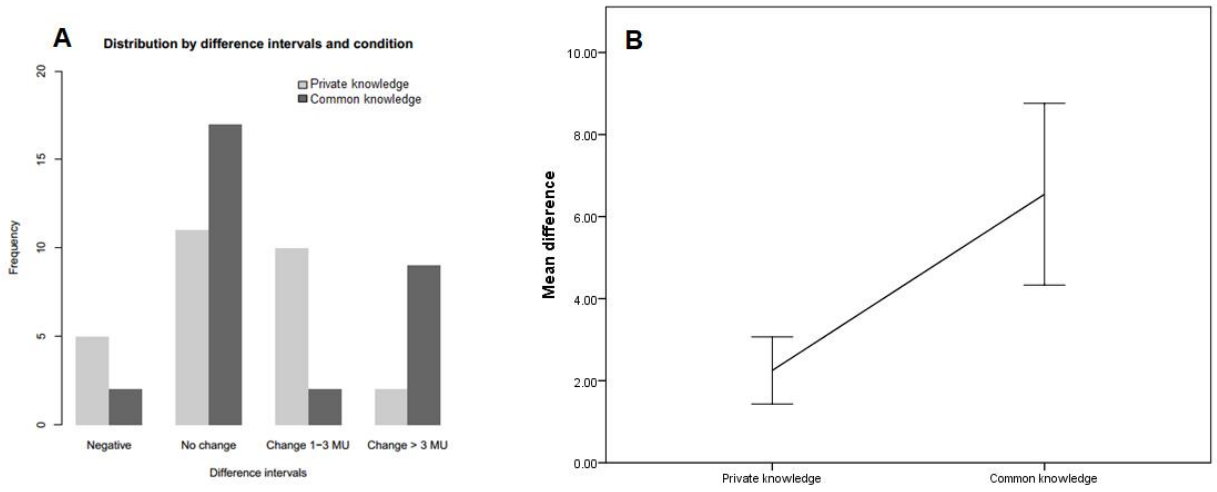


Figure 2.2. (**A**) Distribution of frequencies across categories of change amounts, by condition. (**B**) Mean increase amount by condition, error bars represent 95% CIs.

2.2.2.2. Amount of change

The second prediction we made was related to the absolute amount of change in each condition, to reflect the fact that actors would anticipate that changes known by the observers would raise suspicions about their prosocial intentions. We predicted that only those with very low contributions in the common knowledge condition would find the cost of changing worthwhile, with the hope of sufficiently increasing the probability of being chosen for the additional game. The change amount was calculated by subtracting the initial R3 contribution from the final (changed) R3 contribution, to reflect our expectations that most participants would compete to be chosen for another game by increasing the contribution shown to observers.

We conducted separate Mann-Whitney U tests for the differences in initial contributions between those who chose to increase their contributions and those who did not in the two conditions. We additionally report the results of the sensitivity analyses in brackets after each comparison, i.e. the minimal effect size detectable in terms of the sample size, an alpha level of

.05 and power set at 80 percent. There was no significant difference between the conditions in the sample of participants who chose not to change their contributions (*N*=28, Mann-Whitney *U*=89.000, *p*=.828, d=.08). However, there was a significant difference between those who increased their contributions in the private knowledge condition and those in the common knowledge condition (*N*=23, Mann-Whitney *U*=24.500, *p*=.009, d= 1.258), with participants in the common knowledge condition having lower previous contributions than those in the no knowledge condition. This difference was also reflected in the distribution of the increase between the private and common knowledge conditions, which was statistically significant (*n*=30, Mann-Whitney *U*=179.500, *p*=.003, d=1.241).

2.2.2.3. Partner choice decisions

An exploratory analysis of partner choice based on changed and unchanged contributions in the two conditions showed that change had a marginally positive association with being chosen in the private knowledge condition (Phi=.366, *p*=.053), while this was not the case in the common knowledge condition (Phi=.067, *p*=.713). However, since we were primarily interested in the decisions of the actors being observed, the sample of observers from the original experiment was too small to reach any clear-cut conclusions. To further investigate how change reflects on partner choice, we collected data from additional observers, who also showed a preference for interacting with partners who did not change their initial contributions in the majority of cases, unless the unchanged contribution was very low.

We ran a random-effects logistic regression (using the STATA software) with partner choice as the dependent variable, and dummy variables for (1) the unchanged contributions and (2) the difference between the changed and unchanged contributions as independent variables. The three levels of the unchanged contributions dummy were created to account for three

intervals: low contributions (0 MU - 4 MU), medium contributions (5 MU - 7 MU) used as the base for the model, and high contributions (8 MU – 10 MU). The difference dummy also had three levels tracking our predictions: no difference or a higher unchanged contribution (coded as level 1), a small difference in the interval between 1 and 3 points as level 2 (base), and differences higher than 3, predicted to be costly enough to be competitive (coded as level 3). The dependent variable in the model was the actor with the changed contribution being "chosen", i.e. predicted to have had the higher contribution in R4 out of the pair (coded as 1), as opposed to the actor with the unchanged contribution being chosen (coded as 0). In order to account for individual differences, we also included participants as random intercepts in the model.

The model proved to be significant (N=28; Wald $\chi^2(4)=242.74$; $p<$ 001). Looking at the factors more specifically, the level of unchanged contributions was a significant predictor of whether the changed contribution was picked: when sufficiently low (0-4), the changed contributions were preferred (OR=3.3117; p<.001); when sufficiently high (8-10), the opposite was true (OR=.3407; p<.001); compared to the base. Differences above 3 money units also significantly increased the choice of changed contributions (OR=1.4889; p=.001). The intercept was significant, showing that when the unchanged contributions ranged from 5 to 7, with the difference being small (1-3), unchanged contributions were preferred (OR=0.5996; p=.013). This is to say that when the transfers were the same, participants preferred changed contributions at the low end of the scale, and unchanged contributions at the middle and high ends, which held true when the differences between the two were small, i.e. when the changed contributions were slightly higher. Looking at Figure 2.3, the highest benefit from changing is accrued along the bottom of the horizontal axis, which represents the choices in which the changed contributions were paired with very low unchanged outside options.

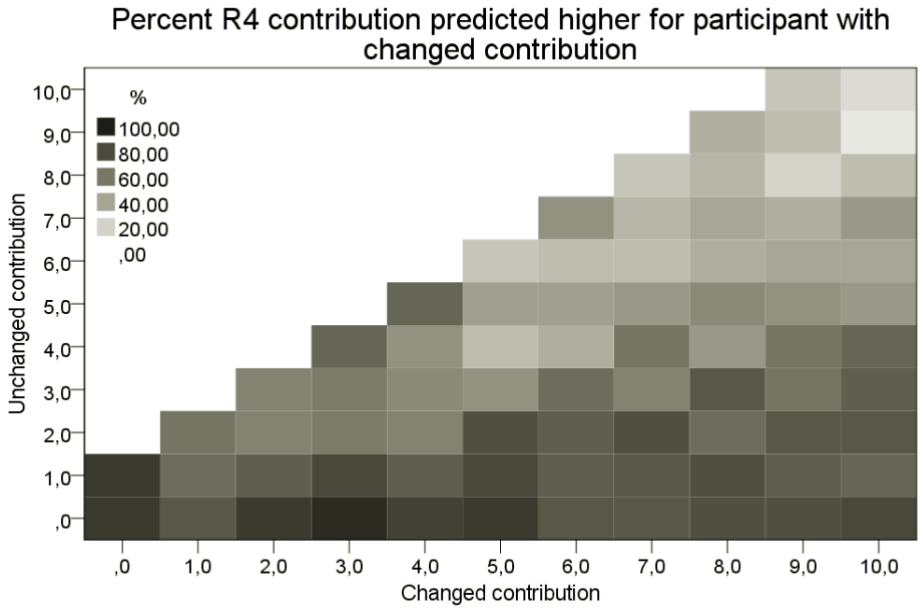


Figure 2.3. Percentages of 'choosing' the actor with a changed contribution (absolute contribution on the x-axis) over an actor with an unchanged contribution (absolute contribution on the y-axis).

2.2.2.4. Social desirability

There were no significant correlations of the SDS-17 scores with either the behaviour in the rounds preceding the information about the observers, or the decisions and amount of change. This held true for both conditions.

2.2.3. Discussion

Maximizing one's profits in a paradigm such as we implemented can be achieved through different strategies: while one is making a (positive) change to increase one's likelihood of getting chosen for a second game, participants could also choose to not engage in impression management (especially in the common knowledge condition) or even decrease their previous contribution in order to maximize their earning from the first game, banking on their partners to give more.

While our first prediction was not confirmed, the observed trend toward higher frequencies of change taken together with the confirmed second prediction that higher increases would be found in the common knowledge condition (because it raises suspicion about the underlying motives) give credence to our hypothesis, i.e. that hidden changes are preferred and that conspicuously incurring the cost to manage impressions of cooperativeness seems to be considered tenable only in cases when initial evidence of prosociality is scarce. If others' impressions of one as a potential partner are already considered by the actor to be negative, such consequent displays of prosociality can seem like a viable option. This can be seen as a 'corrective' reputational strategy, analogous to the experiment by Steele (1975), where an initial negative judgment about one's willingness to cooperate in communal matters increased compliance to later requests for participation in food-sharing projects. In these cases, there might be little to lose in terms of reputation, and possibly something to benefit from if the potential partner's outside options are even less desirable, i.e. if the ratio of genuine cooperators in the population is sufficiently low (Barclay & Reeve, 2012). Furthermore, it is possible that the mere offer to change the contributions nudged some participants to do so in the public knowledge condition as well, despite the fact that it would arouse suspicion in future partners, i.e. that it was due to an experimenter demand effect.

The data obtained from additional observers indicate that the benefits of conspicuous changes seem to be constrained to those instances in which the potential partner's outside options are entering into interactions with evidently uncooperative partners, or instances in which one pays a very high cost to differentiate oneself from others. Large changes in this case could potentially serve as a signal of either an increased willingness to cooperate with the future partner, or contrition for one's previous low contribution. On average, however, observers tend

to prefer partners who did not change their contributions as opposed to those who did, even when the changed contributions are slightly higher. This confirms our third prediction, i.e. it shows that there is an audience discounting effect when attributing prosocial dispositions from observed prosocial choices made in the 'shadow of ulterior motives'.

2.3. Study 2: Evaluations of prosociality in ecological contexts

Prosocial choices in everyday life are, on the surface, different from the straightforward monetary exchanges that can be implemented in economic games. The former often involve small, "mundane" acts of helping (Barclay, 2016), which are difficult to recreate or measure in controlled settings. We chose to additionally address the question of 'audience discounting' by employing a complementary method that further tests whether the pattern of choices of the observers in our lab experiment reflects patterns of choices in our day-to-day social interactions and assess the ecological validity of the results from our lab-based study.

This method also allowed us to enrich the context of the observed interactions and introduce additional relevant factors, as well as collect a larger sample of participants. We hypothesized that audience discounting would be a function not only of changing one's mind, which helps tell apart prosocial motives from those related to impression management, but also of who the audience is. This relates to Hypothesis 3, which states that the effect of audience discounting will be larger if the audience is worth impressing (strategic vigilance aimed at the relevance of the audience). We've touched on what makes an audience worth impressing in Chapter 1: some observers, for instance, one is unlikely to see or interact with again, while other actors might not have the necessary means or status to influence one's desired goals. We refer to this combination of audience features that can produce future benefits as the relevance of the audience. We used vignettes where a decision-maker faces the same payoff stakes as the 'actors'

in the previous study to test the sensitivity of observers' strategic vigilance mechanisms in real life situations with different types of audiences. In other words, the vignettes told stories about protagonists making prosocial choices in everyday settings and changing their mind (or not) after being made aware that an audience would observe their decision. The audience in these vignettes varied according to how much it was worth impressing.

As an example, consider the following scenario with Bob: '*Bob is strolling towards his office building alone. A Red Cross volunteer approaches him to ask for a contribution to a project to help homeless people in the city. The volunteer has coupons for five, ten, and fifty euro. Bob takes out his wallet and takes out a five euro bill. He sees a group of his colleagues returning from lunch, approaching from behind the building and waving. He looks back into his wallet and produces another five euro bill, buys the ten-euro voucher from the volunteer and walks on towards his office headquarters.*' Why did Bob add five euro to his contribution? What if Bob hadn't seen his colleagues, but two pigeons on the sidewalk or a group of tourists, making the same change to finally contribute ten euro? What if he'd taken out another 45 euro and contributed the maximal amount, seen by his colleagues, pigeons or tourists?

In this study, we tested whether the relevance of the audience for the actor or the amount of the increase affects prosocial attributions by asking participants to read such vignettes and then judge the generosity, likability and trustworthiness of their protagonists and make predictions about their future behaviour. Given the above, we predicted that participants will judge protagonists that increase an initial prosocial contribution when noting the presence of a relevant audience as less prosocial, trustworthy and likable than:

P1. protagonists who do not change their initial decision or change it despite no new audience

(derived from H1);

P2. protagonists who change their decision in front of an irrelevant audience (derived from H3).

2.3.1. Method

2.3.1.1. Participants

Our final number of participants amounted to 230 (159 female; *M* age=31.83, *SD*=10.53), collected through the CEU university mailing list and online network dissemination, who filled out the survey on a voluntary basis. We determined the total sample in order to accommodate linear contrasts based on t-tests, expecting a large effect d=.80 with power set at 80% and alpha at .05. This resulted in a minimal sample size of 26 per group. The study was approved by the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

2.3.1.2. Experimental design

We varied the scenarios to include the options where the protagonist made no change, a small change (e.g. 5 euro), or a large change (e.g. 45 euro) to their initial pro-social choice. We also varied the type of audience that the protagonist took note of and who could observe the final decision. We included manipulations of no audience, an irrelevant audience, and a relevant audience. In the no-audience scenarios, the protagonist's attention was shortly averted to an event irrelevant for impression management (e.g. the presence of pigeons). In the irrelevant audience, the protagonist noted the presence of people he/she was not connected to (e.g. unknown tourists). In the relevant audience conditions, the protagonist noted the presence of people that are worth impressing (e.g. their boss).

We did not run the combinations of irrelevant audience and no change, and irrelevant audience and large change, as the critical test of our hypothesis lay in the comparisons of the

three audience levels at the level of a small change (H3). The ensuing seven conditions were ran with three different vignettes describing actors making different types of prosocial choices: giving money to charity (monetary), contributing food to a Christmas party (monetary), and helping a friend move (effort).

After reading a vignette, participants were asked about the generosity and trustworthiness of the story's protagonist and how much they found him or her likable. They were also asked to predict the likelihood of this protagonist making a prosocial choice in other contexts, with no impression management incentives.

2.3.1.3. Procedure

Each participant saw all three vignettes (where different protagonists contributed to charity, a Christmas party, and helped a friend move) presented in random order. For each vignette, they saw a random combination of the two independent variables (type of audience and type of change made by the protagonist).

After reading the vignettes, participants were first asked to answer a comprehension question about the final outcome of the prosocial action (e.g. "How many slices of cake did Gemma buy for the Christmas party?"). If a participant failed the comprehension question of a vignette, we excluded their answers to test questions from the analysis pertaining to the misunderstood vignette (17 in the first story, 18 in the second, 71 in the third). We then asked participants to assess the likelihood that the main protagonist of the story would make specific prosocial choices when IM incentives are absent, on a Likert scale ranging from very unlikely (1) to very likely (5). We included four specific actions: two reflecting prosocial actions of a monetary type, which consisted in buying someone lunch when they've forgotten their wallet

and purchasing an expensive Secret Santa gift, and two where the cost was time or effort, which consisted in proofreading a report and help with carrying a large box. Finally, we elicited ratings of generosity (1) and trustworthiness (2) of the actor in the story, and asked how much participants had liked him/her, measured on a 5-point Likert scale (1 - not at all, 5 - extremely).

2.3.2. Results

In the following sections, we present the main analyses of the generosity, trustworthiness, and affective evaluations. We ran multivariate analyses of variance (MANOVAs) for each vignette separately, with the data split across the change variable to investigate the effect of different audiences which was our primary focus. We report the sensitivity of each main test in brackets, at 80% power and an .05 alpha level (two-tailed).

2.3.2.1. Monetary charitable contributions

At the no change level, there was no difference in the evaluations of generosity, trustworthiness or liking of protagonists in the no audience and the relevant audience conditions ($F(3, 59)=.807$, Pillai's V=.039, $p=.495$).

When the protagonists had made a small change to their contributions, however, there was a significant multivariate effect of the audience ($F(6, 166)=5.188$, Pillai's V=.316, $p<.001$). Tests of between-subjects audience effects showed significant differences in generosity ($F(2, 84)=7.962$, $p=.001$, adjusted $R^2=.139$), trustworthiness ($F(2, 84) =19.173$, $p<.001$, adjusted $R^2=.297$), and liking ($F(2, 84)=8. 294$, $p<.001$, adjusted $R^2=.145$).

Protagonists who made small changes in the presence of a relevant audience were rated as less generous ($M=2.906$, $SD=.777$) than those in the no audience ($M=3.594$, $SD=.615$, $p<.001$) and irrelevant audience ($M=3.357$, $SD=.678$, $p=.036$). Similarly, they were also rated as less

trustworthy (*M*=2.594, *SD*=.756) than those in the no audience (*M*=3.656, *SD*=.745, *p*< 001) and irrelevant audience (*M*=3.250, *SD*=.646, *p*=.001). Liking followed the same patters: participants reported liking protagonists in the no audience (*M*=3.500, *SD*=.842, *p*<.001) and irrelevant audience (*M*=3.179, *SD*=.772, *p*=.054) conditions more than those in the relevant audience condition (*M*=2.625, *SD*=.907). There was no difference in generosity, trustworthiness or liking ratings between irrelevant and no audience conditions.

Finally, even when the change was large, the effect of audience remained present (*F*(3, 63)=10.007, Pillai's V=.323, *p*<.001). The between-subjects effect of the audience was significant for all three dependent variables, i.e. generosity (*F*(1, 65) = 25.898, *p*<.001, adjusted R^2=.285), trustworthiness (*F*(1, 65) =17.581, *p*<.001, adjusted R^2=.201), and liking (*F*(1, 66)=26.447, *p*<.001, adjusted R^2=.278). Protagonists who made big changes were evaluated as less generous (*M*=2.914, *SD*=.612) than those in the no audience condition (*M*=3.838, *SD*=.799, *p*<.001), perceived as less trustorthy (*M*=2.629, *SD*=.598) than the no audience counterparts (*M*=3.297, *SD*=.661, *p*<.001), and liked less (*M*=2.371, *SD*=.877) than the protagonists in the other condition (*M*=3.351, *SD*=.753, *p*<.001).

2.3.2.2. Monetary public good contributions

There was no difference between the no audience and relevant audience conditions in either of the dependent variables at the no change level (*F*(3, 48)=.561, Pillai's V =.034, *p*=.644). At the small change level, there was again a significant effect of the audience (*F*(6, 176)=5.918, Pillai's V =.336, *p*<.001). Tests of between-subjects audience effects showed significant differences in generosity (*F*(2, 89)= 21.550, *p*<.001, adjusted R^2=.311), trustworthiness (*F*(2, 89) =11.673, *p*<.001, adjusted R^2=.190), and liking (*F*(2, 89)=13.692, *p*<.001, adjusted R^2=.218).

Pairwise comparisons showed significant differences in generosity evaluations between the relevant audience condition (*M*=2.917, *SD*=.937) and the irrelevant (*M*=4.188, *SD*=.738, p<.001) and no audience conditions (*M*=3.788, *SD*=.650, *p*<.001), and no difference between the irrelevant and no audience conditions. The same pattern held for trustworthiness evaluations, where the relevant audience condition (*M*=2.806, *SD*=.889) was again different from both the irrelevant audience (*M*=3.813, *SD*=.859, *p*<.001) and the no audience condition (*M*=3.485, *SD*=.755, *p*=.007). Finally, participants reported liking the protagonists in the no audience (*M*=3.606, *SD*=.864, *p*=.002) and irrelevant audience (*M*=3.936, *SD*=.759, *p*<.001) conditions more than those in the relevant audience condition (*M*=2.806, *SD*=.889).

The MANOVA with audience type as predictor was also significant at the big change level (*F*(3,64)=8. 292, Pillai's V=.280, *p*<.001). The effect of the audience was significant for evaluations of generosity (*F*(1, 66)=12.844, *p*=.003, adjusted R^2=.150), trustworthiness (*F*(1, 66) =9.657, *p*=.001, adjusted R^2=.114), and liking (*F*(1, 66)=24.311, *p*<.001, adjusted R^2=.258). Evaluations of generosity were significantly lower for those in the relevant audience (*M*=3.000, SD=.956) than those in the no audience condition (*M*=3.744, SD=.818). Protagonists who performed the large change with no audience were also trusted more (*M*=3.359, *SD*=.843) than those in the relevant audience condition (*M*=2.750, *SD*=.841). Finally, participants reported liking the no-audience protagonists more (*M*=3.513, *SD*=.756) than those who made the change in front of the relevant audience (*M*=2.556, *SD*=.809).

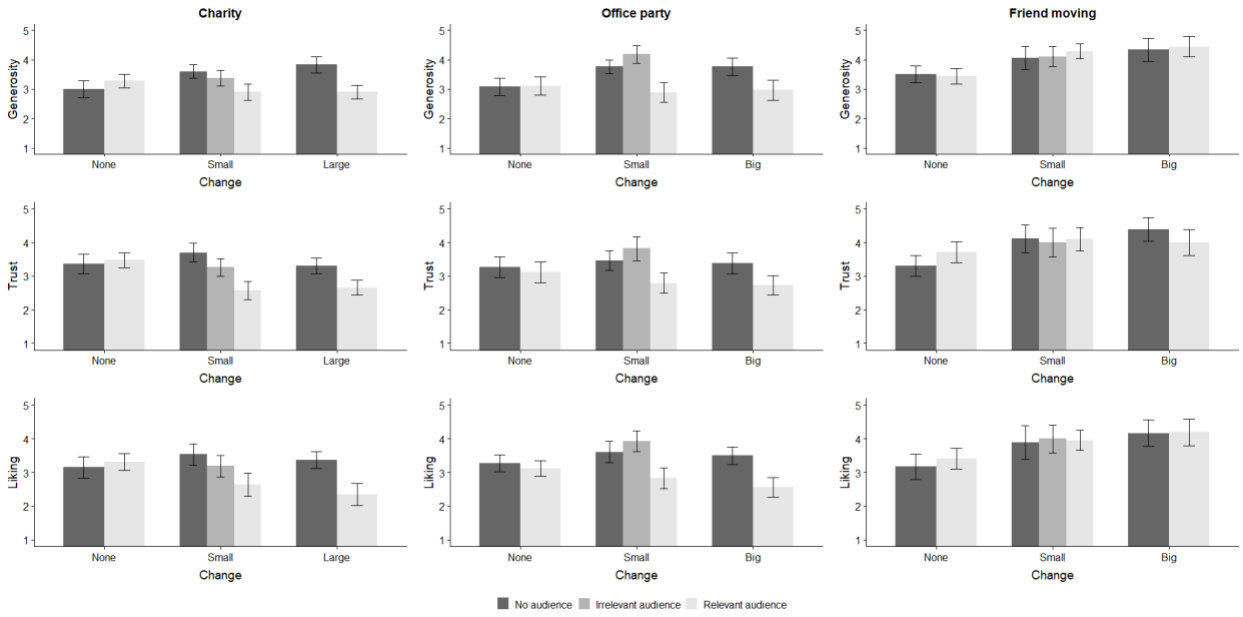


Figure 2.4. Mean evaluations of generosity, trustworthiness and liking shown separately for each vignette, across change and audience conditions (error bars represent 95% CIs).

2.3.2.3. Helping a friend

The results of the third vignette differed from the first two, in that there was no significant effect of audience type at any level of the change variable (no change: $F(3, 64)=2.691$, Pillai's V=.112, $p=.054$; small change: $F(6, 106)=.476$, Pillai's V=.052, $p=.825$; large change: $F(3, 30)=1.603$, Pillai's V=.138, $p=.209$). We ran an exploratory analysis to check whether the change had an impact on the no audience and relevant audience levels. Our analyses showed there was a significant effect of change at the no audience level ($F(6, 136)=3.576$, Pillai's V=.273, $p=.003$). Between-subjects effects showed significant differences in evaluated generosity ($F(2, 69)= 7.040$, $p=.002$, adjusted $R^2=.145$), trustworthiness ($F(2, 69)=12.116$, $p<.001$, adjusted $R^2=.238$), and liking ($F(2, 69)=6.796$, $p=.002$, adjusted $R^2=.140$).

The protagonists who stayed to help their friend for another 5 hours, i.e. until they were done moving ($M=4.257$, $SD=.780$) were regarded as more generous than those who decided not

to stay (*M*=3.487, *SD*=.837, *p*=.002). Trustworthiness was also affected, with protagonists who made no change (*M*=3.297, *SD*=.878) seen as less trustworthy than those who made small (*M*=4.000, *SD*=.849, *p*=.004) or big changes (*M*=4.057, *SD*=.765, *p*<.001). Liking followed the same pattern: participants reported to like those who made no change (*M*=3.162, *SD*=1.093) less than those who made either small (*M*=3.808, *SD*=.939, *p*=.049) or big changes (*M*=3.886, *SD*=.758, *p*=.003).

Change was also significant the relevant audience level ($F(6, 130)=5.328$, Pillai's V=.395, $p<.001$), with significant differences in generosity ($F(2, 66)=17.076$, $p<.001$, adjusted $R^2=.321$) and liking ($F(2, 66)=6.184$, $p=.003$, adjusted $R^2=.132$), but not trustworthiness. Those who made no changes were seen as less generous (*M*=3.485, *SD*=.755) than those who made either a big (*M*=4.200, *SD*=.632, *p*<.001) or a small change (*M*=4.086, *SD*=.658, *p*<.001). Liking was affected in the same way, with the protagonists who had not made a change evaluated as less likable (*M*=3.394, *SD*=.864) than protagonists in the small change (*M*=3.829, *SD*=.618, *p*=.049) or the big change (*M*=4.029, *SD*=.664, *p*=.006) conditions.

2.3.2.4. Prosocial action predictions

Finally, we analyzed the likelihood predictions of actors performing certain specific behaviours by calculating an average of the predicted likelihoods of four specific actions (offering to buy lunch when the person had forgotten their wallet; proofreading a report they will not get credit for; helping carry a large box; buying an expensive Secret Santa present) for each protagonist. We then analyzed the data for each story separately, split across levels of change, using univariate ANOVAs with audience as the independent variable and the average likelihood as the dependent variable.

The predictions of these specific actions closely resembled the evaluations of generosity,

trustworthiness and liking given above. When there was no change, participants rated the protagonists from the no audience and relevant audience similarly in all three stories (there were no differences in the action predictions).

At the small change level, the audience variable had a significant effect on participants' predictions in the charity ($F(2, 84)=5.185$, $p=.008$, adjusted $R^2=.089$) and office party vignettes ($F(2, 89)=11.863$, $p<.001$, adjusted $R^2=.193$), but not in the helping a friend vignette. In the first vignette, protagonists from the no audience condition were rated as more likely ($M=3.500$, $SD=.704$) to help than the protagonists from the relevant audience ($M=2.925$, $SD=.689$, $p=.006$) condition. In the second vignette, those who changed their contributions in front of a relevant audience were also rated as less likely to help ($M=2.868$, $SD=.136$) than those in the no audience ($M=3.692$, $SD=.145$) or the irrelevant ($M=3.714$, $SD=.150$) audience conditions.

Finally, at the big change level, audience again had a significant effect in action predictions in the first two vignettes ($F(1, 65)=15.905$, $p<.001$, adjusted $R^2=.184$; $F(1, 66)=11.691$, $p=.001$, adjusted $R^2=.138$, respectively), but not in the third. In the case of charity contributions, those who contributed in front of a large audience were judged as less likely to help later on ($M=2.734$, $SD=.845$) than those so did so in the no audience condition ($M=3.493$, $SD=.711$, $p<.001$). The same held true for the second vignette, where the protagonist who bought more cake knowing their boss would be there was judged as less likely to help in the future ($M=2.652$, $SD=.126$) than the protagonist who didn't get information about observation ($M=3.616$, $SD=.126$, $p=.001$).

2.3.3. Discussion

Our main analyses show that psychological mechanisms used for evaluating prosociality are sensitive to information about audience value: it is not any public behaviour that is seen as

suspect, but, specifically, public behaviour that can lead to future reputational benefits for the actor. Strategically vigilant participants infer prosocial preferences in view of their attribution of the intention to impress: they understand that the intention to impress is dependent on the type of audience observing a particular action. Prosocial choice is indeed evaluated in such a fine-grained manner, which corresponds to predictions derived from our modulated audience discounting hypothesis.

The results of our third vignette, however, were not significant. It is conceivable that actions which might otherwise be labelled as self-promotion are not perceived as such when they are directed at one's friends, because they are privy to a long history of one's actions. The (post hoc) hypothesis is that a single prosocial choice aimed at people with whom one has had repeated interactions was judged by our participants as too insignificant to alter the audience's inferences about one's prosociality, and thus the participants did not ascribe this change to IM. It is also possible (but less plausible) that investing actual, obvious effort (such as time or physical labour) is perceived in a qualitatively different manner than providing money toward a cause or buying a cake for a Christmas office party.

Finally, the same analysis done on the composites of specific prosocial action predictions followed the same patterns as the evaluations of more abstract, general dispositions like generosity and trustworthiness. This provides additional robustness to the claim that the mechanisms involved in disposition attribution are, at least in part, the same ones used for navigating and predicting everyday social interactions in the cooperation domain.

2.4. General discussion

Our studies dealt with the problem of impression management in contexts where strategic incentives for making a good impression are salient. The first study shows that people's capacity

to manage impressions is powerful enough to anticipate an audience's subtle interpretations of the observed behaviour. More precisely, people know when their intentions to impress will be revealed and have negative effects. Impression management, in this view, incorporates representing others' representations of our intentions to impress. It involves computing third-order meta-representations: my intention is that $_3$[her belief is not that $_2$[I intend to change $_1$[her belief about me]$_1$]$_2$]$_3$. (see also O'Grady, Kliesch, Smith, & Scott-Phillips, 2015, for evidence that higher order mind-reading is easily performed in our day-to-day life).

Our participants found raising suspicion about their intentions to impress to be tenable only when they had very low initial contributions, and changed them to a high degree. They weighed the pros and cons of revealing their intention to impress others and found it worthwhile only when it was beneficial to induce uncertainties about their initial choice, i.e. when it was unambiguously uncooperative. Observers' choices in the same study illustrate that the benefits of impression management are constrained to those instances in which (1) one's alternatives are entering into interactions with uncooperative partners, and (2) instances in which one pays a very high cost to differentiate oneself from one's competition. The decisions of this second batch of observers match the perceptions of primary actors in our common knowledge condition of the original experiment: people are, in fact, sensitive to the likely interpretations of their self-promoting behaviours, which allow them the chance of maximizing their payoffs (observers' outside options notwithstanding).

Our findings add to the literature on the adaptiveness of cognitive mechanisms, allowing us to hypothesize that humans are endowed with a set of rather powerful inferential mechanisms which have the function of both managing our own reputations and forming accurate beliefs about others as reputation managers. Seeing as detecting desirable social partners was one of the

problems our ancestors would have faced in their environment (Cosmides & Tooby, 1992), the resulting selection pressures likely lead to a preference for assorting with those who are, among other things, willing to help and share the benefits of joint ventures with others (Barclay, 2013) – as well as giving rise to complementary selection pressures to *be chosen* for joint ventures, especially by valuable partners. It is thus expected that psychological mechanisms which help to signal one is a suitable cooperator would've evolved (Barclay, 2016), and can be tracked developmentally to children as young as five (Warneken, Sebástian-Enesco, Benjamin, & Pieloch, 2019). On the other hand, humans should also possess mechanisms dedicated to parsing who the valuable partners are from observed choices, such as strategic vigilance (Heintz et al., 2016). Questioning the intentions behind prosocial behaviour is a strategy that can be employed for partner competition on a biological market (Barclay, 2016), but also as a mechanism of partner *choice* – it can be employed to weed out those who are unreliable cooperators. There are plenty of instances where these mechanisms appear to be at work. The growth of online social networking sites has brought the arms race into the open, with their affordance for "cheap" advertising across various domains. In a social environment where strategic motivations are often front and centre, the question of whether and how one can effectively communicate one's prosocial dispositions to vigilant observers is not only of theoretical interest, but also carries practical implications for everyday life.

One way to credibly 'advertise' one's prosociality is by employing indirect strategies when trying to bring one's desirable qualities to others' attention. Packaging self-serving assertions with (irrelevant) negative information (Schlenker & Weigold, 1992), "basking in the reflected glory" of others or distancing oneself from those with whom association could negatively reflect on the self (Cialdini et al., 1976), and creating a context where one is asked

about – rather than asserts – one's qualities (Tal-Or, 2010), are all strategies of self-promotion people have been shown to take advantage of. However, even these strategies aren't always effective, and possibly depend on the audience's reading of the situation. For example, "humble-bragging" (presenting positive information about oneself in the form of a complaint) has been found to reduce likability and competence ratings rather than improve them (Sezer, Gino, & Norton, 2018). Bragging – even subtly – about one's prosocial actions both on- or offline reflects badly on the braggarts in general (Berman et al., 2015) – especially when one has no other information about their reputation.

Study 2 further points to the importance of distinguishing the impact deliberated impression management has on different desirable prosocial traits, namely, generosity and trustworthiness – as well as the type of helping behaviours which are more subject to the adverse effects of self-promotion. While generous actions tailored to relevant audiences impact trait evaluations when the prosocial choice involves giving money to a charity or buying refreshments for an office party (both monetary contributions), the results of the third vignette indicate that an even larger number of contextual variables, such as friendship or the type of helping, needs to be taken into account. It would be interesting to further explore differences between actual and perceived costs of different types of helping behaviours and their commensurability, and how these differences reflect on attributions in the domain of prosociality.

We also show that observation alone does not necessarily lead to a decrease in prosociality evaluations if the observers are regarded as inconsequential to the actor's wellbeing, i.e. if they are people one is unlikely to meet again. Previous studies have mostly focused on either relevant audiences (future partners) or failed to explicitly define audience features, and consequently their value. It is plausible that, in cases where information about the audience is

missing, people will err on the side of caution and act as if the audience is, or will be, relevant at some point in the future. False negatives in this case would have probably been more costly for one's reputation and subsequent loss in valuable partnerships than false positives, pushing psychological mechanisms toward a presumption of importance where observation is concerned (Haselton & Buss, 2000).

Effects of minimal cues of observation under certain conditions (Bateson, Nettle, & Roberts, 2006; Haley & Fessler, 2005) lend credence to this assumption of relevance. However, a recent meta-analysis of the relationship between observability and generosity found a stronger effect of passive observers in engendering generous behaviour as opposed to experimental peers (Bradley, Lawrence, & Ferguson, 2018). Our results also point to the need of more nuanced interpretations of audience effects. One interesting direction for future studies could be disentangling *who* the 'default' or imagined audience is when it comes to minimal observation effects, as well as – crucially – introducing more finely-grained strategic dimensions to the prosocial choices being evaluated.

Chapter 3: The Influence of audience quality on generosity and observers' trust decisions

3.1. Introduction

In the previous two chapters, we outlined the relevance of observation and partner choice for prosocial behaviour and provided some initial empirical evidence about the importance of the proposed relevant audience characteristics which go beyond its mere presence or knowledge about the actor's behaviour. We've also shown that 'third-party' observers take audience relevance into account when making social judgments about actors' generosity and trustworthiness, and that potentially self-interested motives underlying prosocial choice influence self-reported affective reactions of these observers (i.e. their liking of a protagonist in a hypothetical scenario). In this chapter, we further expand our look into the effects of audience characteristics on advertising prosociality on the one hand, and observers' scepticism on the other. While partner choice has been demonstrated to play a role in increasing cooperation above and beyond observation (Barclay & Willer, 2007, Sylwester & Roberts, 2010), various factors which should be taken into account when deciding if an observer is worth 'impressing' have not often been included as variables in experiments investigating reputational concerns in cooperative contexts. In the current study, we addressed the question of one such audience variable, which we hypothesized would have an effect on initial prosocial choices when players were informed of potential future interactions with the audience.

3.1.1. Predictions about the willingness to compete via prosocial choice

The first question we aim to address in this chapter is whether audience quality, operationalized as the difference in the payoffs from interactions with different types of

observers, influences prosocial choice. One possibility is that the willingness to incur a cost in order to be seen as a good cooperator only emerges when that cost is offset by the possible future gains (one could call this strategy "the rational advertiser"). On the other hand, the cognitive mechanisms underlying the target behaviour might not be as finely tuned as to calculate costs and benefits "down to the dollar", and could set impression management in motion as a consequence of the mere eventuality of a future interaction in which some good is going to be divided (no matter its absolute value). Furthermore, social approval combined with the small material gain coming from a low-quality audience (from which one is unlikely to recover the initial prosocial investment) might be enough to offset the cost of the signal. Finally, there is the possibility of mere observation having the same effect, as suggested by eye-cue studies (Haley & Fessler, 2005; Burnham & Hare, 2007), regardless of the actors' beliefs about possible future interactions; or even an effect of an "implicit audience" where the lack of observation is not made explicit and salient. We attempt to disentangle these possibilities in our experiment and explore how varying audience quality affects the decision to advertise cooperativeness.

The various inputs to a mechanism that produces audience effects, and how they interact to subsequently result in impression management strategies, are not straightforward, which could be one of the reasons for the scarcity of studies observed in the literature. For one, the quality of an audience as a group need not be homogenous. It is easy to think of situations in which one or two members of a given observing group carry more weight for an individual's well-being than the rest. As an example, we can take a doctoral thesis defence during which the members of the committee are crucial to convince of one's competence, whereas the occasional student or family member listening in is less instrumental for achieving the relevant goal (and, ideally, doesn't need as much convincing).

Having one such high-value individual in the audience can thus lead to an onlooker's mistaken attributions of the relevance of others, or the audience as a whole. In a similar vein, contributions to common goods which benefit many unrelated individuals can sometimes seem altruistic, but in fact be directed towards the benefits of one or a small number of those affected whose well-being is valued by the actor. As an example, one can consider generous donations to universities made by former alumni, whose children have preferential treatment in the admission process (so-called "legacy admissions") and will benefit from the university's funding in the long run (Jones, & Pittman, 1982). It is thus inevitable to start the investigation of audience quality in what might seem as an over-simplified manner, by considering a rather straightforward inequality about when audience effects can be expected in a prosocial context.

Let *b* denote the benefits one can accrue through future cooperative interactions with one member of the audience, and *n* the size of the audience. Let *c* stand for the cost of the prosocial action afforded by the situation, which could secure those benefits. The use of the word *could* here is crucial, because the success of the action to produce the desired attribution and consequently affect partner choice is not guaranteed and depends on several other factors whose combined influence we shall here, for the sake of clarity, refer to jointly as the probability that the action will lead to future benefits from a given audience member, *p*.

In this case, we can conclude that audience effects will occur when:

$$\sum_{i=1}^{n}(b_i * p_i) > c$$

i.e. when the cost of the action is smaller than the probable future benefits to be had from interactions with the audience members.[10] In the context of this study, we look at the situation where both *n* and *p* are equal to 1 (when there is only one audience member, with whom the actor is certain to interact with in the future) and compare the frequency of prosocial decisions across two experimental conditions (low- and high-quality audience), and two control conditions (passive observer and no mention of observation). In the case of only one observer (who will not be able to communicate what they've seen to relevant others through gossip), the cost of the prosocial action therefore should not exceed the expected probable benefits of cooperating with the audience member. In this chapter, we refer to the observers with which such conditions hold as a 'high-quality audience' (the opposite being a 'low-quality audience').

Taking the above into account, we made the following predictions for the behaviour of the first-movers in our study:

P1. We made the strong prediction that a higher proportion of prosocial decisions would be observed when actors expected the future benefits of interaction with the observer to offset the cost of the prosocial choice than when there were no apparent strategic incentives to behave prosocially (high-quality audience – no audience distinction).

P2. Given the previously outlined results of the link between prosociality and observability, we expected the proportion of prosocial decisions to be lowest when no

[10] This assumes the actor will, in fact, be able to enter into cooperative interactions with all audience members, which we've already pointed out in Chapter 1 might not always be the case. When it is not, *n* should denote the number of audience members with whom the actor thinks it is likely they will be able to interact with: in this sense, *n* more accurately corresponds with the *number of expected interactions* with the audience than its sheer size.

mention of observation was made, as opposed to the conditions in which the existence of observers was known.

Exploratory analyses. As we expected the decisions in the first economic game to be influenced by strategic considerations of appearing cooperative when such advantages were made salient (i.e. in some conditions, but not in others), we looked at whether initial displays of prosociality were less predictive of subsequent cooperative decisions when made with the knowledge of observation and/or possible future gains.

3.1.2. Audience perceptions of prosocial displays in different contexts

Extending the discussion about the influence of contextual cues on the credibility of prosocial displays, this study correspondingly addresses the potential flipside of knowingly advertising to high-quality audiences. In section 1.5., titled *The Catch-22 of Prosocial Impression Management*, we outlined the rationale for the uniqueness of advertising desirable traits in cooperative contexts and the evolutionary basis for audiences being 'sceptical' about self-aggrandizing prosocial displays (Bird & Power, 2015). That public displays of prosociality are taken with a grain of salt and lead to less indirect reciprocity has been experimentally shown (Simpson, & Willer, 2008; Jordan, Hoffman, Nowak, & Rand, 2016), as well as modeled in evolutionary, game-theoretic frameworks (e.g. Hoffman, Yoeli, & Nowak, 2015).

Our experiments from Chapter 2 also addressed perceptions of strategic *changes* in contributions, as well as the influence of the relevance of targeted audience to the actor on subsequent prosociality attributions based upon the changes. The results from these studies indicate that the features of the audience (whether it is one with which an actor can expect future interactions or not) play a part in deciding if a change was made for strategic reasons and if it

will be discounted, decreasing evaluations of prosociality in the cases where the audience is deemed to be relevant.

No such overt cue of advertising was given in the current study; however, observers did receive information about their partner's prior knowledge regarding future interactions when making the first (observed) decision. Because audience quality is a cue similar to audience relevance (in that it might invite scepticism about the motivations underlying prosocial choice), strategically vigilant observers should take it into account when entering into new interactions with observed actors. Hence, we made the following predictions for data gathered from the participants who were given the role of observers in the current study:

P3. Specifically, we predicted that observers would take the context of the prosocial display into account when deciding how much to trust one's partner in a subsequent interaction, such that more trust will be afforded to those who made the cooperative decision without the expectation of subsequent interactions, and with less strategic incentives to act prosocially.

P4. Owing to the same over-arching rationale of this thesis, we also predicted that hypothetical partner preferences (who one would rather trust *fully*, between players from the four conditions) would mirror the above, with participants generally preferring partners who made the cooperative decision in contexts where it is less likely to reflect a strategic decision (most saliently, in the no audience condition) to those with the knowledge of potential future rewards.

P5. Finally, we predicted that cooperative players in all conditions would be trusted more than uncooperative ones, and that this difference might be especially evident

in the cases where the failure to act prosocially was done even with full knowledge of observation and subsequent interactions (i.e. in the high-quality audience condition).

3.2. Method

3.2.1. Participants

We determined the overall sample size according to a power analysis of the audience sample needed for a subsequent ANOVA of the effect of treatment on trust decisions towards *cooperative* dictators to achieve 80% power, and set the cut-off point at N=70 per condition. There were several contingencies for data collection, one of which was an analysis of hypothetical and real choices once a sample of ~70 had been collected, in order to ascertain whether the two decisions differ in a meaningful way.[11]

Consequently, a total of 925 participants completed the study; 463 in the Dictator groups and 462 in the Audience groups. Both the dictators and audience were further divided between 4 conditions: no audience (Dictators N = 106, Audience N = 108), passive audience (Dictators N = 140, Audience N = 140), low-quality audience (Dictators N = 109, Audience N = 108) and high-quality audience (Dictators N = 109, Audience N= 106).

All participants were recruited through Amazon Mturk, an online crowd-sourcing platform, with eligibility criteria set to being proficient in English and being of age (18 or older). Each participant was given a base-pay of $2, and their additional earnings were calculated by pairing the decisions of Dictator and Audience participants in corresponding conditions. The study was approved by the Research Ethics Board at the University of Guelph.

[11] The study was pre-registered and our predictions can be accessed using the following link: https://osf.io/sqgrd

Owing to the nature of online research, we had several criteria for exclusion in order to ensure the economic games were understood, and that the beliefs of the participants reflected those intended in their respective condition. To be included in the final sample, participants had to answer the following questions correctly: the second of two comprehension checks; both manipulation checks (where applicable); and the „bot check“ (to ensure the data collected was from a human). Furthermore, we excluded all participants who indicated they wanted to retract their data from the experiment after the debriefing screen, which contained detailed information about the study and the asymmetric information between the conditions. The unequal group sizes between the Dictator and Audience groups in different conditions reflect the additional number of participants recruited after excluding the data which was not viable.

The final sample thus consisted of 625 participants (365 female; *M* age=34.786): 341 Dictators and 284 in the Audience group.

3.2.2. Procedure

The experiment was programmed using the Qualtrics software for online survey administration. Two groups of participants – which we'll refer to as the Dictators and the Audience – participated in one of four experimental conditions in the between-subjects design. Two economic games were used in the experiment: a Dictator game (DG) and a Trust game (TG). The step-by-step procedure showing the progression of experimental steps can be found in Figure 3.1. Below, we describe the procedure as followed by each of the groups.

Dictator group

All Dictators made two decisions in the same order, (1) the DG in the role of the Dictator, and (2) the TG in the role of the Trustee. They were assigned to one of four conditions, which

varied according to the information they received about the second game prior to making the decision in the DG.

In the no audience condition, participants were not given any information about the subsequent game they would take part in. In the passive audience condition, participants were informed that the decision they made in the Dictator game would be observed by another participant („Player C"), but were not told that they could make money from a subsequent interaction with the observer. In the low-quality audience condition, participants were told both about the observer and their role in the TG, as well as the maximum payoff they could expect to receive from the interaction (40 cents). In the high-quality audience condition, participants were also told about the observer and the subsequent TG, only the maximum payoff was different ($1.60).

Participants were then asked to decide how to split an initial allocation between themselves and another (unidentified) recipient in the DG. Each Dictator got 1 US dollar, which they could decide to either share equally (50 cents to self, 50 cents to recipient), or keep the whole amount to themselves. The Dictator game has been used extensively in the economic literature as a measure of prosocial behaviour or more specifically, generosity (e.g. Haley, & Fessler, 2005). We used neutral language to describe the game in order to avoid framing effects (using Player A and Player B to refer to the dictator and recipient roles, and Player C to the observer where applicable).

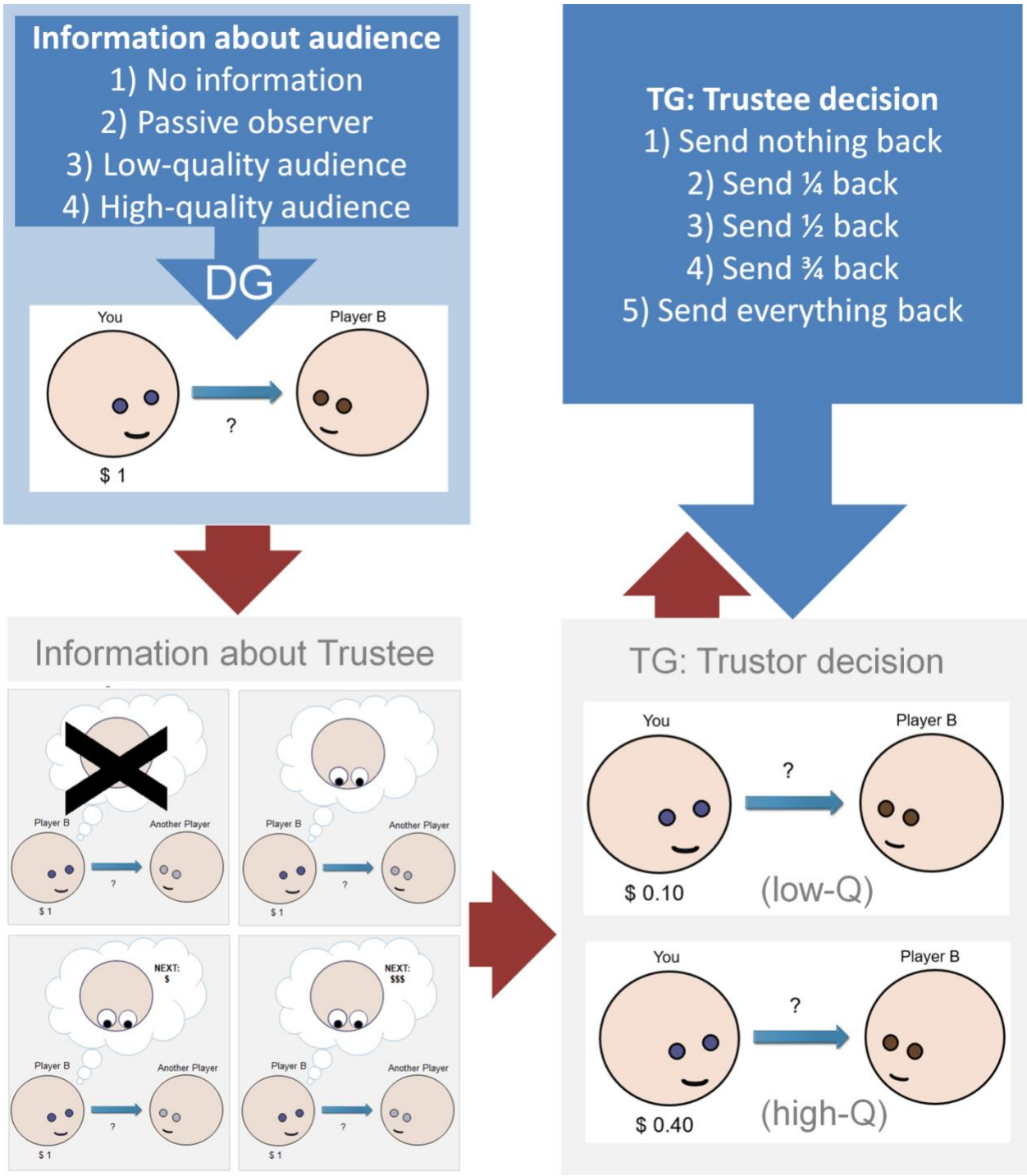


Figure 3.1. Experimental design. Dictator-related manipulations and tasks are in blue, Audience in grey. Red arrows indicate the step-by-step sequence of the two games.

After making their first decision, Dictators were given more exhaustive instructions about their role as Trustees in the TG, and asked to make the second decision. They were told an observer would decide to send a certain amount of money to them, which would be multiplied by four in the process, and that they had to decide which portion of the received sum they would transfer back. Dictators in the high-quality and low-quality audience were always paired with the corresponding observers, whereas those in the no audience and passive audience conditions were randomly assigned to either low-quality or high-quality partners with a 50% chance to be paired with either group. After completing comprehension checks regarding the TG, they were given five options for the back-transfer to choose from: (1) return nothing, (2) return one fourth (what the Trustor had sent), (3) return half, (4) return three fourths (share the profit equally), or (5) return everything. They did not know about the amount they would receive in the transfer at the time of making the decision, as profits from the games were calculated after data from both groups had been collected.

Audience group

The participants in the Audience group took part only in the Trust game in the role of Trustors, which is commonly used to measure trust in economic decisions (Berg at al., 1995). The Trustor is allotted a certain amount of money which they can transfer to the other player, called the Trustee. In this study, there were two types of Trustors, i.e. observers, who belonged to the low-quality or high-quality group, the difference being that the former received 10 cents they could send to the Trustee, and the latter received 40 cents. The money transferred was quadrupled in the process, so the Trustee could receive either a maximum of 40 cents (from the low-quality group) or $1.60 (from the high-quality group). They then had to decide which amount to send back to the Trustor, as explained above. For the Trustor to profit from the

transfer, the Trustee has to return more than a fourth of what they receive, with three fourths being the benchmark for sharing the *profit* equally.

Both the low-quality and high-quality Trustors chose among 11 possible transfer options. Those who received 10 cents could choose to send any amount of cents, including nothing. Those in the high-quality audience who received 40 cents could make their transfer in 4 by 4 cent increments (making up 10 units of 4 cents), also including the option to send nothing.

Before deciding how much money they wanted to transfer, each Audience member was also presented with the decision the Trustee they were paired with („Player B“) had made in the DG (whether they kept the dollar or shared it equally). Additionally, they were also informed of the information Player B had had about the subsequent game (whether they made the choice without any information about being observed; information about being observed but no mention of the TG; or whether they were expecting a low- or high-stakes game to follow).

Counterfactual TG

After making their decision as Trustors in the TG (on which their payment was based), participants in the Audience groups were asked to indicate how much they *would have transferred* to Player B, had they made the opposite decision in the Dictator game. This means that those paired with cooperative Dictators were asked how much they would have sent to an uncooperative Dictator, and vice versa. The answers were indicated on the same scale as in the previous TG (0-10 in 1 cent increments for the low-quality audience, 0-40 in 4 cent increments for the high-quality audience).

General trust level

We included a measure of general trust (Yamagishi & Yamagishi, 1994) at the end of the procedure for the Audience group in order to account for any individual differences in trustfulness in the later analysis of trust towards dictators. The measure consists of six items answered on a 5-point Likert scale ("strongly agree" and "strongly disagree" as end points), with the individual score calculated as the average of the answers to all six items (e.g. "Most people are basically honest.").

Implicit partner ranking

Participants in the Audience group also answered six questions containing "hypothetical partner choice" between two cooperative dictators from different conditions, so that each of the four conditions was compared to the rest. The information provided to Dictators in all four conditions was presented and explained, and Audience members were asked to indicate whom they would rather send 100% of their endowment to in a Trust game, if they'd had the option of choosing (between all possible pairs with regard to condition).

3.3. Results

3.3.1. Dictator group

DG decisions across conditions: Predictions 1 and 2

Our first two predictions concerned the proportions of equal splits in the different conditions, positing that both observation and the expected benefits of the interaction would have a significant impact on the willingness to signal cooperativeness in the Dictator game. The frequencies of choices in the Dictator game can be seen in Table 3.1.

As predicted, the highest proportion of cooperative choices was found in the high-quality audience condition, where 76% of the participants chose to share the dollar with the recipient.

Similarly, 71.4% of those in the passive audience condition also chose the equal split, whereas the lowest proportion of cooperative choice was seen in the low-quality audience condition (36.8%). The decisions in the no audience condition were at chance level.

	No audience	Passive audience	Low-quality audience	High-quality audience	n
Keep dollar.	44 (47.3)	30 (28.6)	25 (36.8)	18 (24.0)	119
Split equally.	49 (52.7)	75 (71.4)	43 (63.2)	57 (76.0)	224
N	93	105	68	75	N = 341

Table 3.1. Frequency of DG choices (percentages inside each condition in parentheses).

We ran a logistic regression model with the decision in the Dictator game as the dependent variable (2 levels: keep dollar, split dollar equally), condition as an indicator categorical predictor with four corresponding levels, and the no audience condition as the reference category. The model was significant (χ^2=12.177, *df*=3; *p*=.007). The intercept was not significant, reflecting the chance-level DG choices in the no audience condition. Further inspection of the estimates showed an increased likelihood of cooperative choices both in the high-quality audience condition (*Wald*=9.396, *df*=1, *p*=.002) and in the passive audience condition (*Wald*=7.282, *df*=1, *p*=.007). Specifically, for cooperative choices, the odds were 2.844 times higher in the high-quality audience condition than in the condition with no mention of observers and 2.245 higher in the condition where observers were mentioned (but no explicit information was given about their role in the future game or a potential payoff from an interaction with them). Dictators in the low-quality audience did not significantly differ in their likelihood of making a fair split from those in the no audience condition.

Decisions in TG: Exploratory analysis of Trustee decisions

Since we expected strategic considerations to influence decisions in the Dictator game in some conditions and not in others, we looked at the associations between choices in the Dictator

game and choices in the Trust game in each condition (Figure 3.2.). All correlations were significant and positive. In other words, acting cooperatively in the DG was associated with higher back-transfer choices in the TG across all conditions.

This association was strongest in the case of the low-quality audience (Kendall's τ=.575, p< .001), with lower associations in the high-quality (Kendall's τ=.354, p=.001) and passive audiences (Kendall's τ=.380, p< .001). The lowest association, though still significant, was found between the DG and TG choices in the no audience condition (Kendall's τ=.215, p=.03).

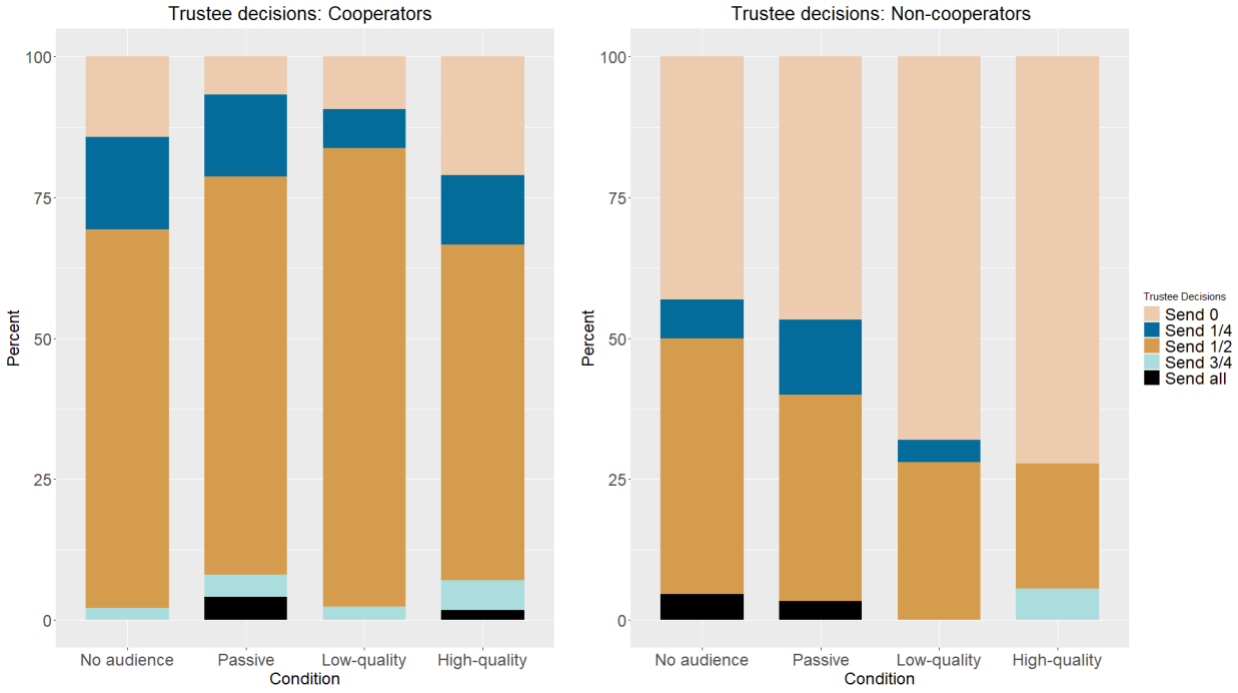


Figure 3.2. Percent of Trustee decisions made by cooperative (DG decision: split dollar) and non-cooperative (DG decision: keep dollar) Dictators, across conditions (N=330).

3.3.2. Audience group

Are cooperators trusted more?

In order to check the basic assumption that cooperative players are trusted more than uncooperative ones, we conducted one-way ANOVAs for each condition separately, with the

percent of the Trustor's allocation as the dependent variable, and the Trustee's decision in the Dictator game as the independent, categorical variable (shared / didn't share the dollar). The results show that cooperators did receive higher transfers (were trusted more) in each of the conditions than non-cooperators (Table 3.2; Figure 3.3.).

		n	*M*	*SE*	*F*	*p*
No audience	Kept $1	29	28.621	6.925	31.260	.000
	Kept $0.50	39	74.872	4.940		
	Total	68	55.147	4.929		
Passive audience	Kept $1	17	20.000	6.751	56.774	.000
	Kept $0.50	48	79.167	4.015		
	Total	65	63.693	4.721		
Low-quality audience	Kept $1	30	16.667	5.325	61.443	.000
	Kept $0.50	48	74.375	4.772		
	Total	78	52.180	4.785		
High-quality audience	Kept $1	22	36.364	8.592	12.065	.001
	Kept $0.50	45	69.556	5.202		
	Total	67	58.657	4.849		

Table 3.2. One-way ANOVA results with percent of TG allocation sent by observer as the dependent variable and Trustee's DG decision as the independent variable, for each condition.

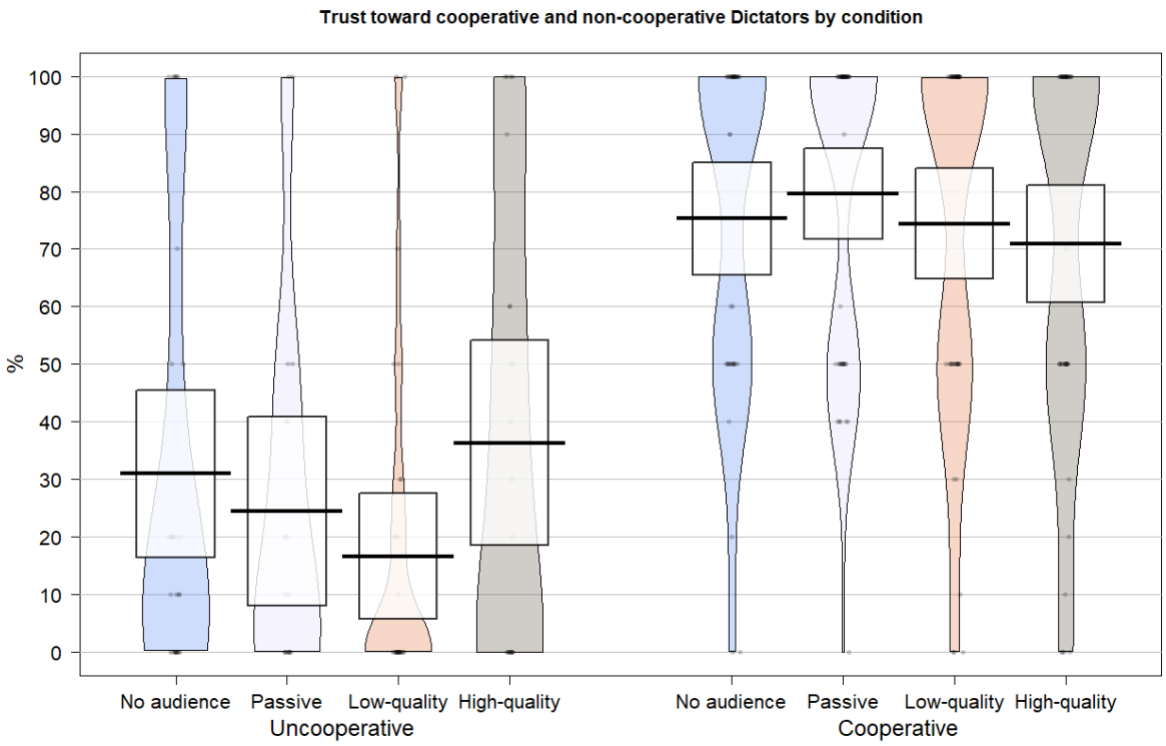


Figure 3.3. RDI plots of transfers to cooperative and uncooperative Trustees. Horizontal lines show means, bands show 95% CIs.

Trust decisions toward cooperative Dictators from different conditions

After running separate independent t-tests to compare the means of actual TG decisions toward a cooperative Trustee and the hypothetical transfer decisions to a cooperative Trustee (from those Trustors who were paired with previously uncooperative Dictators) and finding no significant differences between the hypothetical and real choices, we merged them to perform the analyses of interest (transfers to cooperative Dictators across different conditions).

A univariate ANCOVA with condition as a fixed factor, general trust level as a covariate, and percent of transfer to a cooperative dictator as a dependent variable did not show significant differences in the mean transfers between the conditions (F=2.00, p=.095, see Figure 3.4.).

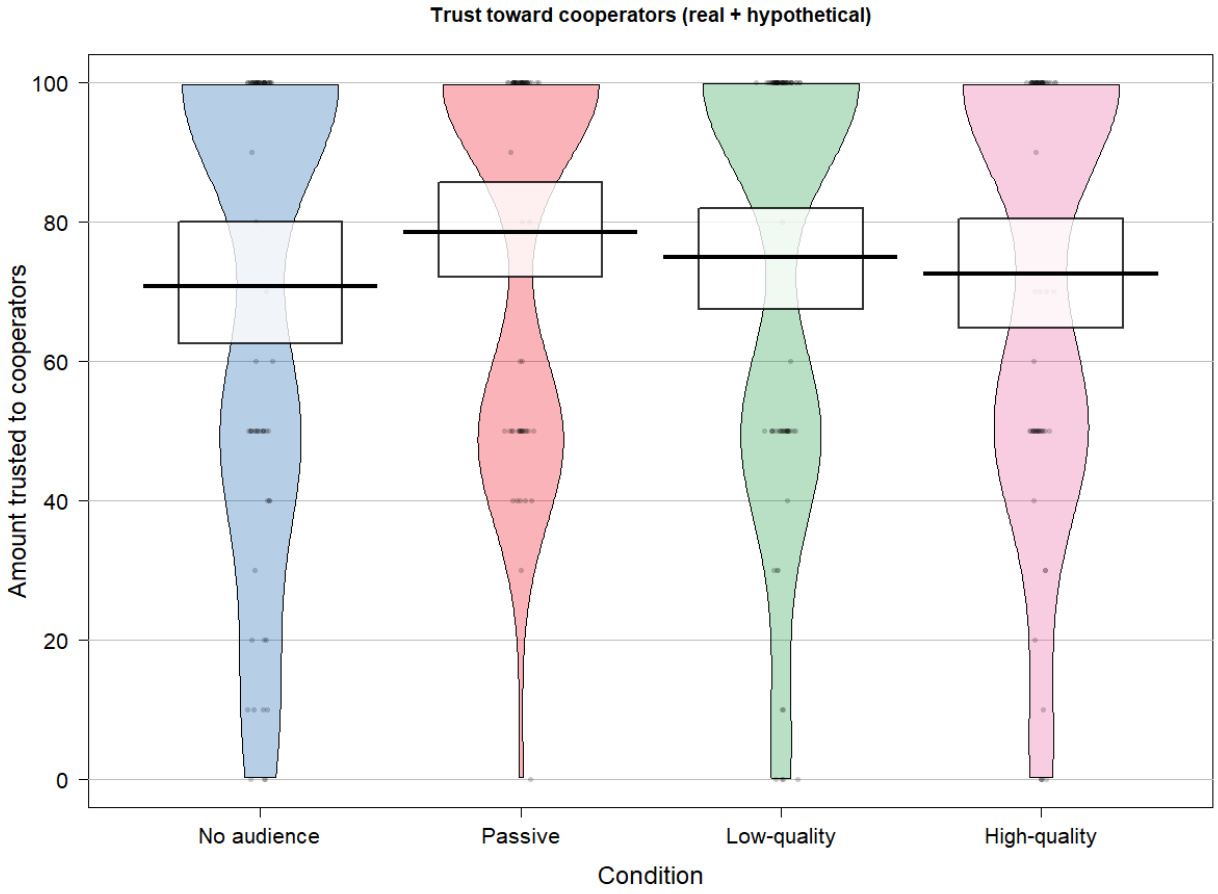


Figure 3.4. RDI plots of transfers to cooperative Trustees. Vertical lines show means, bands show 95% CIs.

Partner preferences

Figure 3.5. shows the frequencies of between-condition partner choices for each presented pair of cooperative dictators. We performed χ^2 goodness-of-fit tests on each pair, which showed significant differences in all cases. Overall, the results of the choices were as we predicted: cooperators from the no audience condition were preferred over those from the high-quality audience condition (χ^2= 8.803, *df*=1, *p*=.003), passive audience condition (χ^2=24.845, *df* =1, *p*<.001, Phi=0.1761) and the low-quality audience (χ^2= 39.563, *df*=1, *p*<.001, Phi=0.373). Those from the passive audience condition were preferred over both low-quality audience

cooperators (χ^2= 27. 268, *df*=1, *p*<.001, Phi=0.308) and high-quality audience cooperators (χ^2=11.042, *df* =1, *p*=.001, Phi=0.1972). The only surprising difference was between the low- and high-quality audience cooperators, in which high-quality cooperators were preferred (χ^2=45.761, *df* = 1, *p*< .001, Phi=0.401).

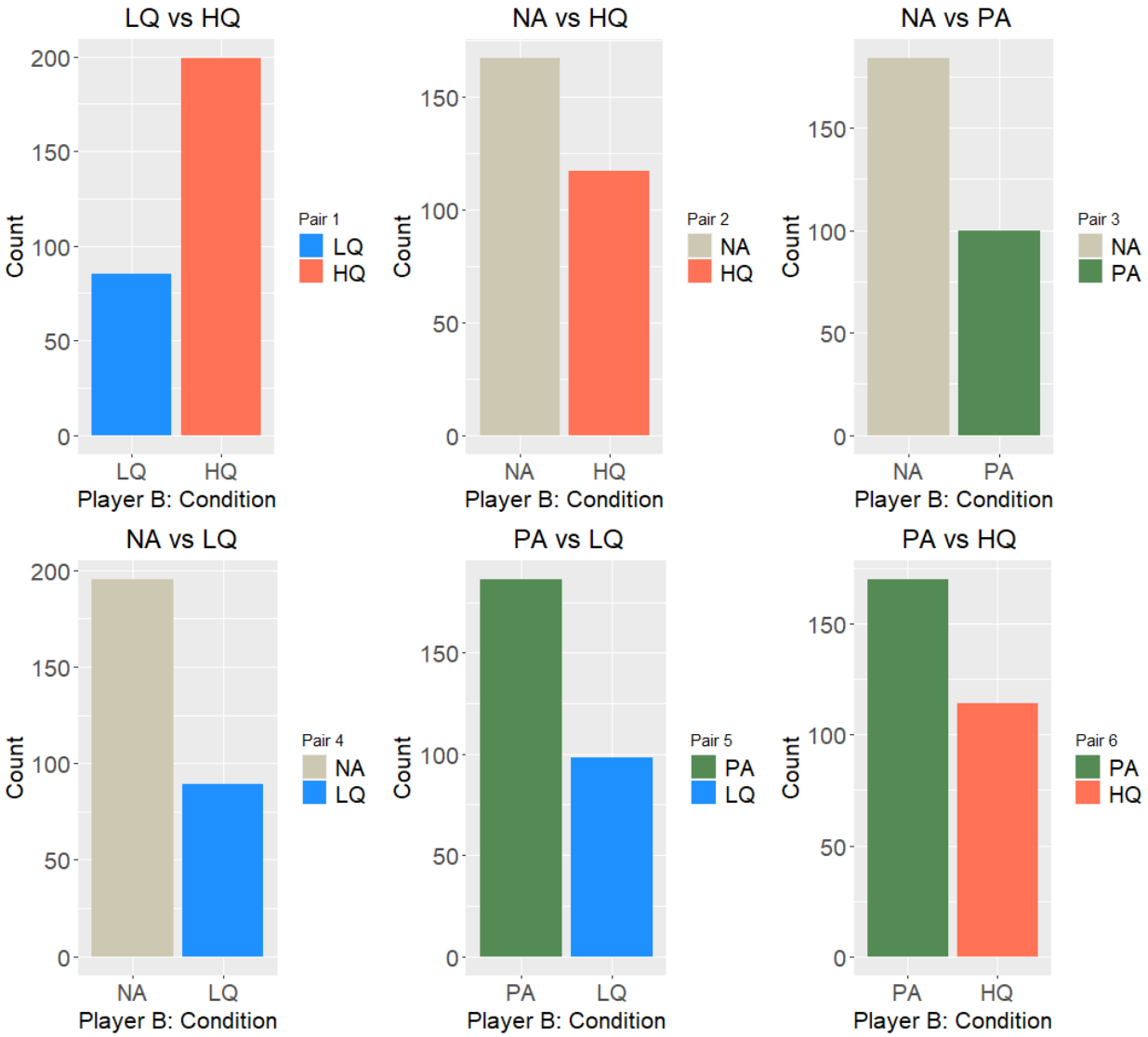


Figure 3.5. Frequency of choosing previously cooperative dictator as the Trustee in a hypothetical game.

3.4. Discussion

The data gathered from the Dictator groups suggests that audience quality, conceptualized as the expected payoff from an interaction with (one of) its members, is a factor which makes prosocial decisions more likely. The cost of this signal seems to be weighed against the benefits of future interactions, such that cooperative signals are produced more frequently in the case when these benefits have the potential of offsetting the cost (our high-quality audience condition) than when no mention of observation is made. On the other hand, making actors aware of an observer in tandem with the fact that they will not be able to profit from advertising cooperativeness (the low-quality audience condition) in the subsequent interaction does not significantly boost the frequency of prosocial choices. Thus, contrary to the additive effects we considered as a possibility – of one's action being observed and reacted to in the second game combining the influence of mere observation with that of a material payoff (however small) – our participants cooperated more readily only at the mention of a *high*-stakes future game. Interestingly, disclosing the low quality of the audience seemed to counteract the effect of observation seen in the condition with passive observers, whose virtual presence also significantly increased cooperation in the Dictator game.

There are several reasons for the low-quality audience to not have caused an increase in cooperation similar to that found in the passive and high-quality audiences. For one, it could be that the cost-benefit analysis leading to the decision to act cooperatively (in those individuals who would normally act selfishly), as given by the simplistic go-no go rule in the beginning, corresponds to the mechanism underlying self-presentational strategies in cooperative situations: that gain is really computed "down to the dollar" and that prosocial dispositions are advertised accordingly. In this sense, because audience quality was the only information relevant to the task

ahead given to the Dictators before making their decision, they were likely to tailor their strategy to this cue of the partner's ability to confer benefits (Barclay, 2016). This is a plausible explanation for finding a difference in effects between the low- and high-quality audiences, but what about the unexpected influence of the passive observer?

The explanation we find most likely is that observational cues as such are by default linked to the presumption of a high-quality, potentially relevant audience – until it is proven differently. As such, the knowledge about there being a passive observer would have prompted the same reaction as the information about the high-quality observer due to an error-management-like mechanism (Buss & Haselton, 2000; Haselton & Galperin, 2012) which would take all observers as potentially being the latter. A more similar, and trivial reason (that should nevertheless be mentioned) would be an experimenter demand effect, or more broadly, the context of the experimental situation. Dictators in the passive audience condition might have believed that there was a reason for the observation to have been mentioned, and thus acted according to this belief. However, this wouldn't necessarily bring our finding into question, as they would have still been reacting on the presumption that the passive observer would be able to confer benefits to them in a later interaction, or was important for the rest of the experiment in another way.

When it comes to the results of the Trust game, we successfully replicated previous findings which show that cooperators are afforded more trust and preferred as partners in cooperative interactions than non-cooperators (e.g. Gambetta & Przepiorka, 2014). We did not find an effect of the strategic incentives behind the cooperative decision on the audience's trust toward dictators in different conditions. However, our predictions were mostly borne out of the forced-choice tasks in which the whole of one's endowment was at stake, except for the

interesting finding that participants preferred cooperators from the high-quality audience (who had the most incentive to signal deceptively) over low-quality audience ones. One of the contributing factors might have also been that the Trustors perceived it likelier for those expecting high rewards to be willing to share more of the "pie" in the back-transfer than those who expected only small benefits (even with the maximum of which they would still be at a loss).

Chapter 4: Framing effects reveal differences in attitudes towards social rules

4.1. Introduction

The empirical studies we've presented in the chapters so far have been either conceptually or explicitly related to social exchange: we've looked at how audience features such as an observer's knowledge about the actor's incentives, audience relevance and audience 'quality' affect people's willingness to manage impressions of prosociality and, on the other hand, how audiences perceive such prosocial displays when observed in potentially self-serving contexts. In this chapter, we move away from investigating more-or-less explicit economic transfers and focus on a related phenomenon: that of framing effects and self-image concerns in hypothetical reports about social rule-breaking.

In Chapter 1, the effects discussed as potential outputs of an evolved impression management mechanism included conformity and, more pointedly, the difference between compliance to and acceptance of group beliefs and attitudes (Sowden et al., 2018); as well as self-enhancement phenomena (Brown, 1986; Krueger, 1998). In the following study, we try to distinguish between exogenously imposed rules which could be seen as resulting in *compliance* (rather than *acceptance*); in other words, those which are neither internalized nor perceived as relevant social norms (thus being 'fair game' to break); and those rules which are relevant to participants' self-image, and by the same token less likely to be influenced by external incentives or contextual change. We further investigate how impression management motives might influence these responses via two methods in different, locally relevant scenarios which vary in the degrees of perceived normativity.

4.1.1. Motivations underlying rule abidance

The motives underlying abidance to social rules can relate to concerns about one's own welfare, about the welfare of others, about self-image, social image, and likely a myriad others. Why, for instance, does one pay for a tram ticket? It might simply be in order to avoid getting fined. It can also be because of genuine concern about public goods and a willingness to financially contribute to the maintenance of the tram lines they often use. Yet another option is that they think they are not someone who would cheat on public transport: they are averse to free-riding because it would go against what they (would like to) think of themselves (*self-image*) and what they would like others to think of them (*social image* and consequently *reputation*).

The latter two motives - constructing or maintaining a positive self-image – can make us 'feel good' about ourselves and proud when we are able to signal that we are moral people to others (e.g. Bénabou & Tirole, 2002). Crucially, the motivation to maintain one's moral self-concept explains why people are likely to cheat if the prospect of being punished is null, but do so only a little, in ways that spare them the need to update their self-evaluations of how moral they are (Mazar, Amir, & Ariely, 2008). Self-image concern is a likely motive underlying social rule abidance that is both pervasive across cultures and grounded in evolved psychological traits for impression and reputation management (Heintz, Karabegovic, & Molnar, 2016).

One empirical challenge, for social scientists, consists in revealing which motives influence choices to abide (or not) by a given social rule. In the case of paying one's tram ticket, a researcher might ask whether people abide by this rule because they want to be good community members and believe this is required of them to fulfil that role, or if other motives, such as avoiding punishment, are more important. Which rules are followed because of self-

image concerns? Which are contingent on being observed? Which other motives might relate to rule abidance? This empirical challenge is difficult because most attitudes towards rules will involve several types of motives at the same time, which are often difficult to disentangle. In this paper, we use an experimental design that suggests that self-image concerns often motivate rule following, more so for some social rules than for others.

4.1.2. Social desirability and self-enhancement biases as indices of an evolved impression management mechanism

Our method for revealing whether self-image concerns play a role in rule abidance consists in documenting self-enhancement biases, which consist in presenting the self in a positive way, as *more* moral or *more* competent, relative to an imagined 'average' peer (Alicke, 1985). As previously noted, people tend to evaluate themselves as possessing higher levels of desirable traits than others when making social judgments, while the opposite is true when it comes to undesirable traits (Brown, 1986; Furnham, 1986). The same holds true in self-reports of hypothetical (un)ethical actions (Ariely, 2012). In particular, people are sensitive to nuances of these scenarios and their answers are influenced by the possibility of a priori and post-hoc justifications of the actions (e.g. Bersoff, 1999; Shalvi, Gino, Barkan, & Ayal, 2015).

Self-image concerns and socially desirable responding are often viewed as a cumbersome methodological nuisance in self-report questionnaires, especially those dealing with sensitive topics (Krumpal, 2013), because they prevent researchers from getting unbiased estimates of the behaviours or beliefs in question and contribute to the larger issues related to the experimenter effect. Interestingly (though not surprisingly), observers have also been found to make a difference in this domain – for instance, Aquilino, Wright and Supple (2000) showed that the presence of parents (but not siblings; or romantic partners for older participants) influenced

disclosures about alcohol and tobacco use in paper-and-pen surveys. Aquilino (1997) had previously outlined several factors which influence socially desirable responding, one of them being is the probability of negative consequences from the disclosure which corresponds to what we'd predict from an evolutionary impression management perspective. Furthermore, socially desirable responding isn't unique to psychological studies – for instance, it's also been found to influence participants' preferences (see e.g., Kuran, 1987, on preference falsification).

A variety of methods have been designed to overcome social desirability in responding to gather more accurate information (e.g. the randomized response technique, Boruch, 1971; the unmatched count technique, Dalton, Wimbush, & Daily, 1994). One of them includes asking questions indirectly (in the third person), and has been shown to mitigate social desirability when it comes to socially sensitive topics (Fisher, 1993). Experimental economists, on the other hand, have often used material incentives with the hope of crowding-out confounds and getting closer to what people would do in 'real life' (for a discussion on the benefits and constraints of using monetary incentives, see Read, 2005).

In this study, rather than trying to do away with self-enhancement and social desirability, we focused on the insights they can provide by using both self- and other-frames in our hypothetical scenarios of ethical behaviour, as well as combining these frames with an incentivized coordination game and a non-incentivized run-of-the-mill online survey. Our main assumption in doing this was that self-enhancement biases would influence answers in the self-referential frame, but less so in the third-person frame (similarly to reports on cheating by Ariely's (2012) golfers), and that this self-framing effect would be modulated by the normativity of the particular social rule as well as, potentially, the method. Since these biases carry information about the respondents' perceptions of how consequential the issue is for their (social

or) self-image, we believe it is worth exploring them in a more systematic manner, and documenting their effects in view of theoretically relevant variables such as community standards.

4.1.3. Perceived pervasiveness of rule-breaking, rule normativity and context influence rule abidance

Relevant beliefs concerning social rules include the beliefs about whether others abide by them or not, and the beliefs about whether others think that we *should* abide by them. These two beliefs roughly correspond to the concepts of descriptive and injunctive norms (Cialdini, Reno, & Kallgren, 1990), or empirical and normative expectations in Bicchieri's vernacular (2005). They're important because they are strongly related to motivations for abiding by a given rule, above and beyond the fear of being punished. While they often overlap and thus direct behaviour in predictable directions, discrepancies between what is (apparently) 'practiced' and what is 'preached' have resulted in counterintuitive findings, with behaviour seemingly more affected by descriptive rather than injunctive norms and so-framed behavioural interventions (e.g. Corral-Verdugo, Frias-Armenta, Pérez-Urias, Orduña-Cabrera, & Espinoza-Gallego, 2002; Cialdini et al., 2006; Smith et al., 2012), and with empirical expectations also trumping normative ones when in conflict (Xiao & Bicchieri, 2009). Furthermore, these beliefs vary across cultures and differently affect the motivation to abide by a given descriptive norm (Gelfand & Harrington, 2015).

For documenting these beliefs in the current study, we relied on two methods. We used the Krupka-Weber (2013) method to elicit beliefs about participants' perceptions of whether others think a given rule *should* be abided by (what we refer to in the rest of this chapter as *social acceptability*). Secondly, we used a framing method applied to an incentivized pure coordination

game to document beliefs about whether people would actually report their willingness to abide by the rule or not. More precisely, we randomly paired participants and asked them to choose the same option as their partner between abiding by the rule and not abiding by the rule. While the task and the incentive remain the same – being rewarded for choosing the same as one's partner – the type of rule used in the frame was varied. Finally, we also used the same question without the monetary incentive to further investigate the effect of the specific rules and their normativity on socially desirable responding.

Behaviour in economic games is well-known to be sensitive to framing (Alekseev, Charness, & Gneezy, 2017; Bermúdez, 2020). Simply giving different names to equally presented information ('Wall Street Game' vs. 'Community Game') influences the rate of cooperation (Liberman, Samuels, & Ross, 2004). Sometimes the super-imposed context can be unintentional – for example, Ensminger (2004) found that her Orma participants had spontaneously associated the Public goods game with *harambee*, a Kenyan social institution of fundraising to help schools and public projects in their villages. Contextualizing games in view of cultural concepts has shown promise in producing effects - both in the cultures of their origin (Cronk, 2007; Lesorogol, 2007), but also with Western participants who had only been familiarized with the concepts during an experiment (Cronk & Wasielewski, 2008). In our study, we use cultural scenarios as labels that can enable coordination. Crucially, we also vary whether each scenario is presented with a self-referential frame or a third-person frame, i.e. we document self-framing effects on beliefs about rule abidance in the given context.

We predict that the self-framing effect will reflect the strength of self-image concerns for rule abidance in a given scenario. In other words, the biases documented in framing effects will be a means to reveal whether self-image concerns play a role in abiding by the rule in question.

We test this prediction by analyzing framing effects in view of another proxy that influences self-image concerns (and impression management motivations): beliefs about whether it is acceptable to break the rule or not. We expect these two to be related, as self-image concerns are largely influenced by what we believe others expect of us (e.g. Heintz et al, 2015). This is to say, if – in spite of the material incentives of our experiment – we find higher levels of reported rule abidance in the I-frame than in the they-frame, we can conclude that self-image concerns are, at least partly, at work in the motivation to follow the rule.

Another related goal of this study is to show that self-image concerns do not produce a blanket effect on the bulk of what might *a priori* be considered unethical. In fact, we predict that these concerns will vary depending on the local intuitions about acceptability coming from one's expected audience (community), and that documenting biases in self-reports can provide information about what kind of impression people are locally motivated to maintain and advertise.

4.2. Method

4.2.1. Participants

All participants were recruited in Madurai (Tamil Nadu, India), the main site of the study. For the purpose of this chapter, we analyzed data gathered from three separate samples: 230 students who participated in a lab-based coordination game (77 female, 153 male; *M* age=19.97), 101 who participated in an online survey, which was also carried out on site and in the same college (19 female, 78 male, 4 NA; age data missing, but comparable to the other two samples). A further 82 participants were recruited for the Krupka-Weber measure which was done in a

paper-pencil method (30 female, 51 male, 1 NA, *M* age=18.77).[12] The study had previously been approved by EPKEB in Hungary.

4.2.2. Measures and procedure

4.2.2.1. Social norm and self-other frames

We constructed nine short scenarios in the participants' native language (Tamil) with the help of an ethnographer familiar with the local context, in which the protagonist could choose to make either an 'ethical' or 'unethical' choice, i.e. they could choose to act in accordance with a social norm or break it. Participants were tasked to pick one of the actions (ethical/unethical) for each scenario, with the consequences of the choice dependent on the method used (see descriptions of the coordination game and survey below). The nine scenarios ranged from economic transgressions to more culturally relevant dilemmas involving inter-caste marriage or religious (in)tolerance and differed in perceived social acceptability and perceived pervasiveness of their infractions (how likely one thought other people would choose the dishonest option). For example, in the case of bribery, the scenario participants read was the following: *Raj is traveling from Madurai to Chennai by night train. His tickets are not confirmed, it is still on the waiting list. He gets into the train on the waiting list hoping to get confirmed. TTE is around verifying tickets. It is possible to offer money to the TTE and get the ticket confirmed. What will Raj do?* We also included a neutral frame which consisted in the choice between two differently shaded square shapes.

Additionally, questions about the actions in all scenarios were presented in either the third-person singular, like in the above example with Raj as the protagonist, or in the first person

[12] The data was collected as part of the 'Beliefs fostering dishonesty' project funded by the CEU UWI grant – we do not analyze the full set of the data in this chapter, but only the subset relevant to the hypothesis about framing effects in ethical choice.

(asking participants what would *they* would do in the given situation). Each participant saw all nine scenarios in either the self-referential frame or the third-person frame to avoid confusion or potential spill-over effects between the conditions.

4.2.2.2. Coordinating via social norms: economic game

We used a simple matching coordination game where the goal is to anticipate the partner's choices (Mehta et al., 1994), with the above-mentioned scenarios as the main test of beliefs about social norms in everyday behaviour. We incentivized participants to guess an unidentified partner's answer to hypothetical decisions between allegedly ethical and unethical (but potentially socially acceptable) actions. The choice set always consisted of two possible actions – one of them ethical and one unethical. In the above example of verifying train tickets, participants would thus pick between bribing the train conductor or not bribing the train conductor. Their payment in each round of the coordination game depended on whether the choice they picked was the same as their partner's – every matched answer was worth 25 INR, whereas not matching meant participants did not make any additional money in that round.

Each participant was presented with all nine scenarios during their session in the computer lab at their college, with a total of seven rounds of the coordination game for each scenario.[13] The experiment was programmed in Z-Tree (Fischbacher, 2007) such that participants made their choices between the options on the screen. The scenarios themselves were announced by the experimenter and presented in a booklet in Tamil. They played in randomly grouped teams of five – their partner was chosen from the same team, with one

[13] We analyze data from the first rounds of the game only, as we expect it to be the most indicative of both initial beliefs about the pervasiveness of the given behaviour and most likely to be affected by self-image concerns. The dynamics of changing beliefs throughout the rounds of the coordination game will be analyzed and presented as a separate paper.

randomly assigned participant 'sitting out' in each round. For the purposes of this paper, we only look at decisions made in the first round.

4.2.2.3. Survey

The survey was conducted in the same pool of participants, however, it was carried out online. It contained the same questions and format of answers like the coordination game, only participants' payoffs were not dependent on another's answer. In short, they replied to the survey as they would to an ordinary questionnaire, reporting what they themselves or another person would do in a given situation (act ethically or not).

We conducted the survey for two reasons. The first was to explore whether a non-incentivized method (in either the first or third person) would differ from the decisions made in the game, especially in the direction of socially desirable responding in the first person. The second question we wanted to address was an added effect of mentalizing (if any) in the coordination game, i.e. whether participants would anticipate others' self-image concerns when responding to scenarios in which choosing the socially less desirable action is seen as unacceptable.

4.2.2.4. Measure of social acceptability

Finally, we used the measure developed by Krupka and Weber (2013) to elicit judgments about injunctive norms. Mirroring their method, we presented participants with the abovementioned third-person scenarios which ended with the actor choosing the unethical action. We then asked participants to rate how socially appropriate the action was, on a 4-point Likert scale (1 - very socially inappropriate, 2 - somewhat socially inappropriate, 3 - somewhat socially appropriate, 4 - very socially appropriate). Participants were instructed to choose the answer they thought would be most commonly chosen in their group (i.e. aligned with the mode of the

distribution). In this chapter, we primarily use the results obtained from this measure to classify the scenarios with regard to social acceptability and investigate whether these differences reflect on the choices in the coordination game and the survey, for the first- and third- person frames.

4.2.3. Study design and analysis plan

We employed a mixed experimental design for our main research question, with the frame of reference (self/other) as the between-subjects factor and the scenario as the within-subjects factor. We further employed two distinct methods, an incentivized coordination game on site and a non-incentivized survey carried out online. This difference in methods is an additional between-subjects factor in a further analysis focusing on the role of incentives in crowding out impression management motives.

In the following section, we first refer to the results of the Krupka-Weber measure to classify the scenarios according to social acceptability. We then look at the results of mixed-effects models taking into account method, framing and social acceptability. Finally, we present scenario-specific analyses of the framing differences in the coordination game and survey data.

4.3. Results

4.3.1. Social acceptability of unethical choices across scenarios

We classified the unethical choices in our scenarios following the injunctive norm elicitation method (Krupka & Weber, 2013) into three categories: (1) Socially acceptable when the modal response skewed toward the acceptable end; (2) Socially unacceptable, when the modal response skewed toward being unacceptable, and (3) Socially ambiguous when there was no consensus on either side of the scale (bi-modal or uniform distributions). Our scenarios, in decreasing order of social acceptability, were ranked as such: bribing the train ticket examiner, educating one's son instead of one's daughter, income certificate – all predominantly judged to be

socially acceptable; and keeping a lost wallet, overbilling a customer, discriminating on the grounds of religion, which were deemed to be socially unacceptable. Three scenarios showed no clear pattern in terms of social acceptability: littering, recusing oneself from a committee because of a conflict of interest, and supporting an inter-caste vs. a love marriage.

For the purpose of the main analyses, we took the subset of the data that the majority of participants agreed was either socially acceptable overall, or socially unacceptable overall (in other words, we excluded the results of the ambiguous scenarios) to test the influence of social acceptability on the willingness to coordinate on dishonest choices, and check for a possible interaction with the reference frame. We repeated these analyses separately for the survey and game data. A closer inspection of the specific scenarios can be found at the end of the section.

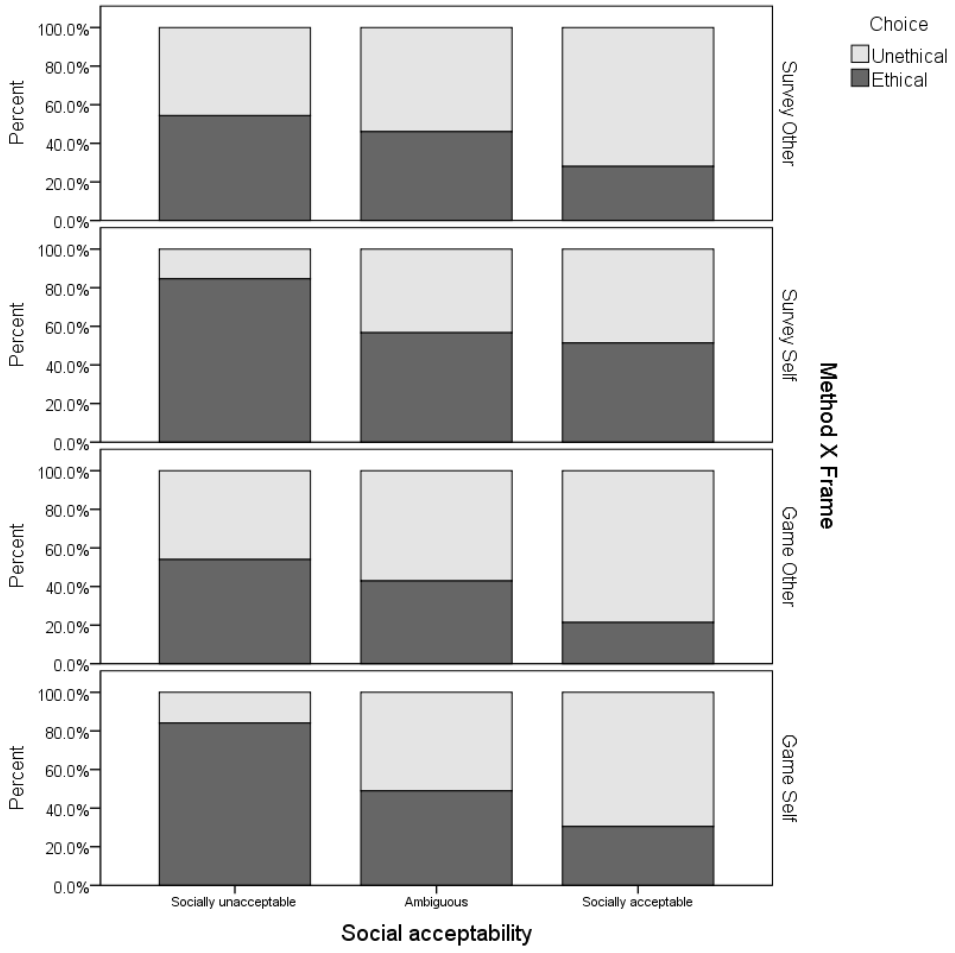


Figure 4.1. Pooled percentages of ethical and unethical choices across categories of social acceptability, presented for each combination of frame and method.

4.3.2. Models

4.3.2.1. The influence of referential frame, method and social acceptability on choice

We first analyzed the aggregate observations from scenarios which were classified as either acceptable or unacceptable, leaving out the ambiguous scenarios, in order to check whether there was an effect of acceptability on participants' choices, and whether this effect interacted with the other independent variables (the framing and the method). We ran a generalized linear mixed model with a logit link using the lme4 R package (Bates, Mächler, Bolker, & Walker, 2015), with choice as the dependent variable (coded as: 1 – unethical, 2 -

ethical). The predictors included in this first model were: acceptability (coded as: 0 – socially unacceptable, 1 – socially acceptable), method (coded as: 0 – survey, 1 - game) and referential frame (0 – other, 1 - self), while participants were included as a random effect to reflect the repeated measures design. We also included two-way interaction terms for combinations between all three main effects, as well as their three-way interaction.

Model	Full model with 2- and 3-way interaction
Fixed	Log-Odds (SE)
Intercept (*base: other-reference, survey, socially unacceptable*)	0.087 (0.070)
Reference (*self*)	0.852 (0.101)***
Method (*coordination game*)	-0.248 (0.09911)*
Acceptability (*socially acceptable*)	-1.267 (0.091)***
Reference x Acceptability	-0.412 (0.122)***
Reference x Method	-0.144 (0.140)
Acceptability x Method	-0.309 (0.122)*
Reference x Acceptability x Method	-0.194 (0.17287)
Random	
ID var	0.313
Observations, Participants	N=1697, N=329
Model Fit	AIC= 1957.8 BIC= 2006.7 logLik= -969.9 deviance= 1939.8

Table 4.1. Generalized linear mixed model of stated ethical choice fit by maximum likelihood (Laplace Approximation), family: logit. Significance levels: * p < .05 ** p < .01 *** p < .001.

All three main effects were significant (see Table 4.1.). The self-frame positively affected the likelihood of making the ethical choice as opposed to the other-frame (*OR*=2.34, *SE*=0.236, *z*=8.46, *p*<.001, 95% CIs [1.92, 2.86]). Ethical choices were less likely in the coordination game (*OR*=0.780, *SE*=0.0773, *z*=-2.50, *p*=0.012, 95% CIs [0.643, 0.948]). Similarly, they were also less likely in scenarios which were classified as socially acceptable (*OR*=0.282, *SE*=0.026, *p*<.001, z=-14.0, 95% CIs [0.236, 0.337]). The 3-way interaction wasn't significant, however, two of the 2-way interaction terms were.

Firstly, there was a significant interaction between frame and social acceptability (*OR*=0.662, *SE*=0.081, *z*=-3.37, *p*< .001, 95% CIs [0.521, 0.842]), which was driven by choices

made in the self-frame, which influenced socially acceptable choices in a different way than it did in the other-frame, especially when taking into consideration the method as well. The second significant interaction was between acceptability and method (*OR*=0.734, *SE*=0.090, *z*=-2.53, *p*=0.011, 95% CIs [0.578, 0.933]), which suggests that participants were more likely to choose the socially acceptable unethical actions in the game than in the survey. In order to gain a better understanding of these interactions and how they relate to the framing effect, we split the data by frame and investigated the dynamics of acceptability and method for each referential frame separately.

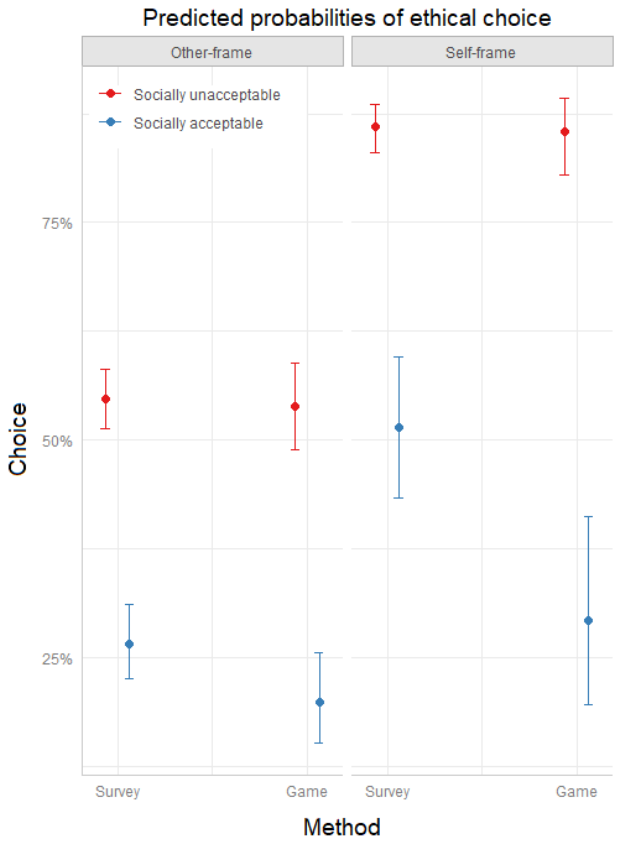


Figure 4.2. Predicted probabilities of ethical choice from the full model.

4.3.2.2. The influence of method and social acceptability on self- and other-frame data subsets

Both subsets (self- and other-frame data) were re-analyzed using similar generalized linear mixed models with a logit link to predict ethical choice, with acceptability, method and their interaction as fixed effects and participants as the random effect. Table 4.2. shows the results from two models for each subset, those with and without an interaction term. The model with the interaction was a better fit for the self-frame data than the one which didn't include the interaction according to the AIC criterion (Table 4.2.) or a likelihood ratio comparison of the two (χ^2=6.7, *df*=1, *p* =0.010). The opposite was true for the other-frame data, where the model without the interaction performed better according to the AIC and BIC criteria as shown in the same table, as well as a likelihood ratio comparison (χ^2= 0.9263, *df*=1, *p*=.336).

Acceptability is a significant main effect in both cases – in the scenarios coded as socially acceptable, the probability of ethical choices is significantly less likely than in scenarios coded as socially unacceptable, in both the self- (*OR*=0.215, *SE*=0.028, *z*=-12.031, *p*<.001, 95% CIs [0.167, 0.276]) and other-frame (*OR*=0.355, *SE*=0.043, *z*=-7.641, *p*<.001, 95% CIs [0.280, 0.449]). This is the only significant main effect in the other-frame subset of the data. In the self-frame data, method also seems to play a part with coordination game choices more likely to be unethical (*OR*=0.709, *SE*=0.093, *z*=-2.621, *p*=0.009, 95% CIs [0.548, 0.917]). Importantly, we again find a significant interaction between acceptability and method (*OR*=0.644, *SE*=0.109, *z*=-2.610, *p*=0.009, 95% CIs [0.463, 0.896]). Looking at the predicted probabilites of the self-frame model (Figure 4.3.A), the method does not influence the high probability of ethical choices in the socially unacceptable scenarios, however it does in the acceptable scenarios: there is a higher number of self-reported ethical choices in the survey than in the coordination game in these scenarios, whereas the same effect isn't found in the other-frame (Figure 4.3.B).

	Self-frame		Other-frame	
Model	No interaction	**Interaction**	**No interaction**	Interaction
Fixed	Log-Odds (s.e.)	Log-Odds (s.e.)	Log-Odds (s.e.)	Log-Odds (s.e.)
Intercept ()	0.712*** (0.095)	0.680*** (0.095)	-0.525*** (0.108)	-0.523*** (0.107)
Method	-0.436*** (0.12848)	-0.344** (0.131)	-0.133 (0.150)	-0.148 (0.150)
Acceptability	-1.609*** (0.127)	-1.537** (0.128)	-1.036*** (0.121)	-0.989*** (0.129)
Method x Acceptability	-	-0.440** (0.169)	-	-0.171 (0.177)
Random				
Participant variance	0.227	0.240	0.387	0.383
Observations, Participants	N=917, N=176	N=917, N=176	N=780, N=153	N=780, N=153
Model Fit	AIC=1002.7 BIC=1021.9 logLik =-497.3 deviance =994.7	AIC= 998.0 BIC= 1022.1 logLik= -494.0 deviance=988.0	AIC=960.4 BIC=979.0 logLik= -476.2 deviance=952.4	AIC=961.5 BIC=984.8 logLik=-475.7 deviance=951.5

Table 4.2. Generalized linear mixed models of stated ethical choice fit by maximum likelihood (Laplace Approximation), family: logit. Bolded titles represent the models with the best fit. Significance levels: * p < .05 ** p < .01 *** p < .001.

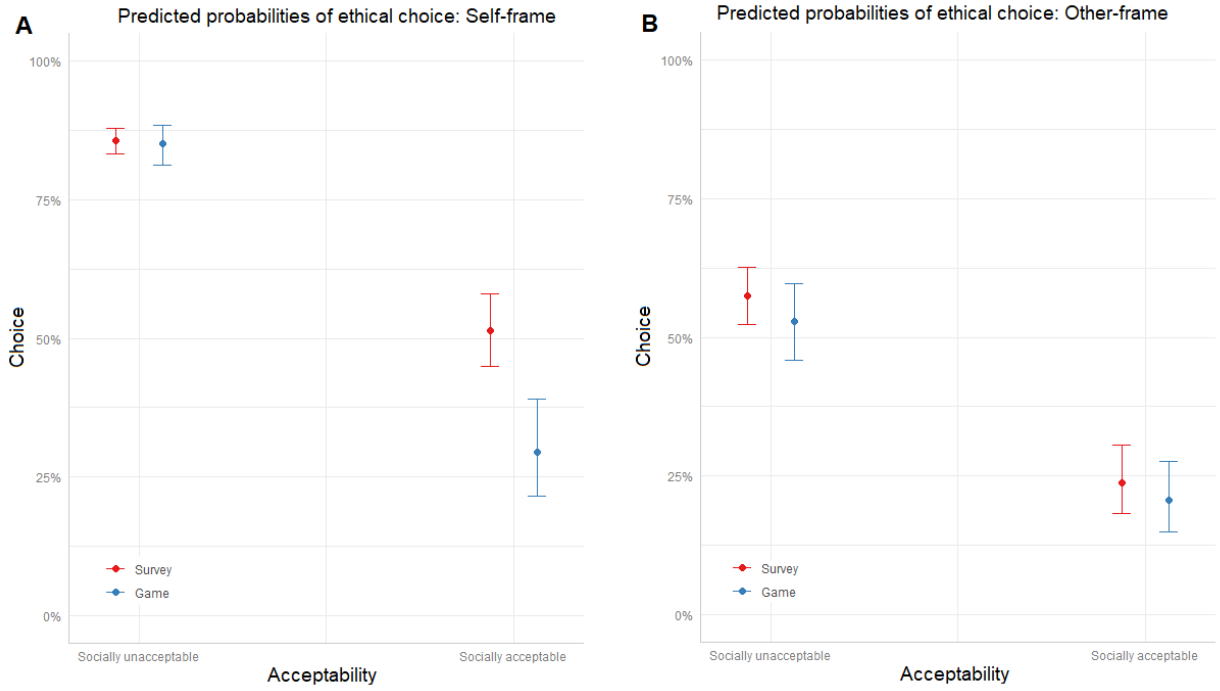


Figure 4.3. Predicted probabilities of ethical choice: (**A**) Self-frame subset of the data, model with an interaction; and (**B**) Other-frame subset of the data, model with no interaction.

4.3.3. Self-other framing differences across specific scenarios and methods

4.3.3.1. Socially acceptable scenarios

We analysed the data from the coordination game and survey separately, using Chi-square tests of independence, with referential frame and choice as the two grouping variables. As making the unethical choice in this case was considered to be socially acceptable, we predicted smaller (or no) differences between the self- and other- frames, especially in the game where participants' payoffs depended on coordination with the partner.

In the train ticket scenario, there was no significant difference between the choices made in the self- and other- frames, in either the game, or the survey. In the gender scenario, the difference was marginally significant in the game (N=184, df=1, χ^2=3.797, χ^2=.051, Phi=0.144), and reached significance in the survey (N=99, df=1, χ^2=4.586, p=.032, Phi=0.215), with more participants reporting they would educate their daughter over their son, if asked in the first person. Finally, in the the income certificate scenario, the differences between self- and other-frames were significant for both methods: participants in the coordination game opted for the ethical choice more frequently in the self-frame than the other-frame (N=184, df=1, χ^2=8.013, p=.007, Phi=0.209), as did those in the survey (N=99, df=1, χ^2=15.705, p<.001, Phi=0.398).

4.3.3.2. Socially unacceptable scenarios

We analysed the data from the socially unacceptable scenarios in the same way as above, with Chi-square tests and grouping on referential framing and choice. We expected significant differences between the self- and other- frames, in both the game and the survey, as we hypothesized that self-bias would be especially strong in those cases where one can expect social disapproval from behaving in a way that is seen as unacceptable.

In the scenario with returning a lost wallet, there was a significant difference between the choices made in self- and other- frames in the game (N=184, df=1, χ^2=6.769, p=.015, Phi=0.192), with more participants reporting they would return the wallet in the self-frame. The same difference in the survey was marginally significant (N=99, df=1, χ^2=4.194, p=.057, Phi=0.206). In the overbilling scenario, the proportion of ethical choices in the self-framing was higher in both the game (N=184, df=1, χ^2=12.267, p=.001, Phi=0.258) and the survey (N=99, df=1, χ^2=9.121, p=.003, Phi=0.304). Finally, the effect seemed to be most pronounced in the religion scenario, where the differences were again significant in both the game (N=184, df=1, χ^2=50.301, p<.001, Phi=0.523) and the survey (N=99, df=1, χ^2=22.425, p<.001, Phi=0.476).

4.3.3.3. Ambiguous scenarios

In the coordination game data, we find no clear preferences for coordinating on either choice in the neutral scenario. However, there was a significant difference in the preference for one of the squares in the survey (N=99, df=1, χ^2=11.073, p=.001, Phi=0.334).[14] There was no significant difference in the conflict of interest scenario, in either the survey or the game. In the marriage scenario, there was a self-other difference only in the survey (N=99, df=1, χ^2=7.106, p=.013, Phi=0.268), but not in the game. In the littering scenario, we find the self-other difference in both the survey (N=99, df=1, χ^2=16.638, p<.001, Phi=0.41) and the game (N=184, df=1, χ^2=5.697, p=.012, Phi=0.176).

4.4. Discussion

In this chapter, instead of trying to account for biases in self-reports, we used them to investigate attitudes and beliefs towards a set of different social rules. We document a self-framing effect towards more ethical choices when scenarios are presented in a self-referential

[14] We don't have a ready explanation for this result, which seems to have been driven by choices in the Survey-Other combination of independent variables. However, since we do not use the netural frame as a baseline in our analyses as was first intended, we believe this should not affect the interpretation of the results from other scenarios.

manner as opposed to indirectly, i.e. in the third person – we find this effect in a non-incentivized questionnaire replicating previous findings on socially desirable responses (e.g. Fisher, 1993). This self-other difference in the likelihood of ethical choices, however, persists even in the incentivized coordination game, though the effect is smaller.

Social acceptability of breaking the specific rules also influences behaviour in both the survey and the game, in both the self- and other-framed scenarios. This perceived social acceptability of (not) abiding by a particular rule also qualifies the self-other effect, together with the method by which it is investigated. Our data shows a decrease of self-enhancing choices in the socially acceptable scenarios when self-referential frames are used in a coordination game as opposed to a non-incentivized survey with the same questions (partly exonerating the experimental economists who vouch for the effects of monetary incentives), while no such difference exists for the socially unacceptable scenarios.

There are several possible explanations for this interaction. Firstly, it is possible that the monetary incentive to coordinate with others nudges people to be more 'honest' about their imagined transgressions and decreases the importance of self-enhancement motives in the cases where the behaviour is perceived to be acceptable by one's community. The coordination game might also provide a plausible justification for the apparently unethical choice (the material incentive), thus making it less reflective of participants' morality than an answer in the survey where no such justification is afforded (Shalvi, Gino, Barkan, & Ayal, 2015). This would be especially true for the ambiguous and socially acceptable unethical choices which might not be internalized as values and would not be seen as informative for the updating of one's moral self-concept (Gino, Norton, & Weber, 2016).

It is also plausible (even if less likely) that the monetary incentive leads people to pay more attention to their beliefs about others' beliefs and enhances perspective taking – thus pushing the ratio of ethical and unethical choices closer to that seen in the same third-person scenarios because of the motivation to coordinate with others (as opposed to replying from the first-person about what *they themselves* would do). This type of process would lead to more unethical choices when the norms are ambiguous or the behaviour is seen as socially acceptable because of a better understanding of the game, in which the main task is in fact to guess what others are more likely to do, rather than choose the option which best fits one's personally held beliefs. The difference in the ratios could then signify a shift towards 'truer' beliefs about the majority's attitude towards the norm.

Inspecting the data from the various scenarios separately is interesting for several reasons. For one, the differences between survey and game responses might provide additional insights about attitudes towards the norms in question, even if they are deemed socially acceptable to some degree. While the game can be a cue as to which rules people are likely to 'admit to' breaking because they aren't necessarily socially sanctioned in their particular community, survey answers could be used to complement this insight with the information of which of these rules are either nevertheless considered reflective of one's moral self-concept or seen as not necessarily socially *unacceptable*, but desirable – to a large enough degree to lead to a motivation to manage impressions.

Relatedly, these framing effects could be complemented with manipulations reflecting the hypothesized emotional bases of rule abidance; shame and guilt. Bicchieri (2005) makes a point to separate social norms (which are conditional on others' expectations) from what she refers to as *moral* norms, which are followed regardless of social context. Elster (2007) similarly

differentiates between social norms, which are observation-dependent and have their basis in (anticipated) shame, and moral norms, which are context-independent and trigger guilt when broken. He also includes a third type, quasi-moral norms, into his taxonomy, which depends on one seeing others comply to a given rule, rather than the opposite (for a comparison of the two approaches, see Dubreuil, & Grégoire, 2013).

Shame is likely an important proximal mechanism which motivates rule abidance, when it is deemed socially unacceptable to break said rules. It's been shown to closely track expected audience devaluation across cultures (Sznycer et al., 2013) and is triggered even when the action itself is not necessarily morally wrong, but could be perceived as such by others (Robertson, Sznycer, Delton, Tooby, & Cosmides, 2018). It would be interesting to further investigate these emotional bases of the self-framing effects we find, especially as they relate to the differences in self-reports of socially acceptable unethical actions between surveys and coordination games.

Chapter 5: Do rule origins affect rule abidance in an economic experiment?

5.1. Introduction

Most of the work in this thesis has relied on the presumption that there is an intuitive threshold for what is seen as prosocial or not, or ethical or not: in other words, a norm or rule against which potential partners can compare observed actions and make their judgments accordingly. However, what constitutes prosociality in a lab setting is difficult to intuit for participants (as well as, sometimes, researchers). We made use of this ambiguity in the following two chapters to investigate two different factors related to (more-or-less) 'artificially' chosen rules and participants' willingness to abide by the same.

Prosocial rules, as we define them, represent instructions about how individuals should behave in a given situation, and are especially relevant when there are conflicts of interest, i.e. when what is good for everyone isn't neatly aligned with self-interest: they have the function of curbing undesirable behaviours in order to increase social welfare. As we already mentioned in the previous chapter, attitudes toward rules can vary depending on the underlying motivations and can thus be affected by a number of different contingencies. One such factor are the beliefs about how many others abide by the rule and whether these 'others' expect one to abide by the rule (Bicchieri, 2005), as well as anticipated emotions of others finding out whether one has broken the rule (e.g. Elster, 2007, see also Sznycer et al., 2013). In this chapter, we refer to yet another factor which should influence rule abidance: that of its origin. Specifically, we look at how rule types affect choices to contribute to a public good in an experimental setting – where the rules are externally imposed (but not sanctioned, as has been done by Gallier, 2020). Our goal is to analyze whether knowledge of the historical processes through which a rule was

specified influences the rate of rule abidance. If people are rational decision makers who strive to maximize future material benefits, they should not be sensitive to this history (as it has no direct consequences on said benefits). However, rule origins might contribute to rule abidance if they provide information about the legitimacy of the rule, which in turn provides information whether potential cooperators will expect one to follow the rule. In other words, they would provide information relevant to calculating whether following the rule is worth the cost in a partner choice ecology.

Rules can can have many different origins, both in terms of the motivations for their abidance, as we've touched on in the last chapter, as well as where (people think) they come from. Legal rules which are proscribed by institutions and often rely on punishment, moral rules (including religious rules) which appeal to internalized values, those which represent what people normally do in a given situation (descriptive norms) or what ought to be done (injunctive norms; Cialdini, Reno, & Kallgren, 1990), and so on. In this chapter, we look at two distinct types of rule origins: democratically chosen rules and rules imposed by a 'rule-maker' or leader. In the latter case, we investigate a similar effect as we did when it comes to prosocial choice: namely, whether scepticism about the leader's intentions comes into play when deciding whether to follow the rule or not.

Several dimensions should make democratic rules 'special'. For one, participating in the process of choosing the rule should give a sense of agency to participants in the rule-making process and increase their willingness to abide by the chosen rule in this way (e.g. Ostrom & Nagendra, 2006; Gallier, 2020). Furthermore – and importantly for the context of this thesis – democratically chosen rules represent others' expectations about how one should behave in a

given situation and provide information about what the majority of one's "audience" considers to be desirable in the context in which the rule is to be implemented.

Any chosen rule as such can be a coordination device, especially in contexts where what is expected or socially desirable is unclear: rules can provide a clue about others' expectations and become a salient coordination tool. However, not all rules are likely to be successful. For instance, the intentions of those making the rules might have an impact in how the rule is evaluated: whether it is seen as being chosen to increase group welfare, or to benefit a specific individual. This effect could be relevant to democratically chosen rules, as well – while they can provide information about the values of one's audience and be perceived as an honest expression of audience values, championing costly rules such as 'contribute everything' could also be seen as a strategic, self-serving choice by audience members. Though this inference is possible, it is much more likely in the cases where one knows about a rule-maker's previous self-serving decision which goes against common welfare. The contrast of a person 'preaching what they themselves haven't practiced' should thus lead to vigilance, in a similar way as ambiguously motivated prosocial actions do (e.g. perhaps the leader wants everyone to contribute the maximum so they can give nothing and maximize their earnings).

In the following experiment, we address this question and look at how different rule origins affect initial rule abidance (our main test), but also how rule-following and subsequent cooperation evolve over repeated rounds. Namely, how do the intentions of those deciding which social rules apply in an experimental situation affect the willingness to comply with said rules? Do democratically chosen rules lead to more rule-following? Finally, when the expectations stemming from an announced prosocial rule are not met, does cooperation suffer more than when no such rule existed to begin with? Does the breaking of *some* rules adversely affect cooperation

more than that of others? Our main hypothesis, based on a strategically vigilant reading of the inferences which can be mad based on the rule-making procedure and the leader's intention, was that the intentions of those setting the rule would be taken into account, in a way similar to partner-choice paradigms. We further predicted that democratically chosen rules would have a larger effect on initial rule abidance, on account of giving information about audience values and their expectations.

5.2. Method

5.2.1. Participants

The sample consisted of CEU students and those registered to the SONA online system for participant recruitment at the Central European University who signed up for the study, with proficiency in English being a prerequisite for participation. The final sample we analyze consisted of 198 participants (98 female, 47 male, 3 other, *M* age=25.84, questionnaire data missing for 21 participants due to technical error): 42 in the Random Leader condition, 54 in the Democratic Vote, and 36 and 66 in the Generous Leader and Selfish Leader conditions, respectively.[15] Sessions lasted 20-30 minutes, depending on condition and number of participants, and consisted of 6-18 players at a time. The experiment was approved by the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

[15] We started collection of the data hoping we could rely on the prosociality of the elected leaders to choose the rule relevant for our comparisons of the different conditions. This contingency did not work as planned, so we altered the design to include preset rules and changed the design of the generous/selfish leader conditions to include the communal game. We thus excluded those sessions in which the chosen rule did not correspond to 100% contributions and which did not include the initial game (6 sessions in total, N=51). However, note that the results of rule preferences include the votes from the Democratic Vote session in which the majority-chosen rule was not 100% (N=12).

5.2.2. Measures

We used a 3-person Public goods game (PGG) to measure rule abidance, and subsequent cooperation within groups of participants. The procedure of the partner-PGG we used is as follows: All participants in the session were randomly assigned to groups of three, which were stable throughout the experiment (i.e. they were composed of the same members throughout the 10 rounds of the game, to the full knowledge of the participants). In each round, every member of the group was given an initial endowment of 20 monetary units (MUs), any whole number of which they could invest into the common box. The number of MUs invested in the common box was multiplied by 2, and then distributed equally between the three group members.

5.2.3. Procedure

Experimental sessions were held at the computer labs of the Central European University. Experiments were programmed and administered using the Z-Tree software for economic game experiments (Fischbacher, 2007). Participants made their decisions in real time, at computer stations in the lab which were partitioned off on all three sides in order to provide anonymity. They were given identifying numbers, which were later used to match the individuals to their earnings in the game and provide payments. The show-up fee for participation was set at 400 HUF, which was added to the total, recalculated earnings from the game.

After consenting to participate in the study, participants were given instructions about the Public goods game. The instructions were presented on the computer screens and simultaneously read aloud by the experimenter in charge of the session, who prompted the participants to ask clarification questions after the instruction phase was over and an example of an imaginary round was presented. A three-part comprehension check with detailed explanations and feedback then followed to make sure everyone understood how earnings would be calculated. Every part of the

check consisted of an imaginary round, which required answers about each player's earnings, and was followed by feedback in the form of the correct answers and an explanation of how they were calculated. After the comprehension checks, we presented participants with five possible "rules" referring to the amount of contributions one should invest in the common box in the upcoming game. The same five rules were presented to groups in all conditions, in the same order:

1. In each round, contribute 20 MUs, your whole endowment, to the common box.
2. In each round, contribute at least 15 MUs to the common box.
3. In each round, contribute exactly 10 MUs of your endowment to the common box.
4. Never contribute to the common box.
5. In each round, contribute whatever you want to the common box.

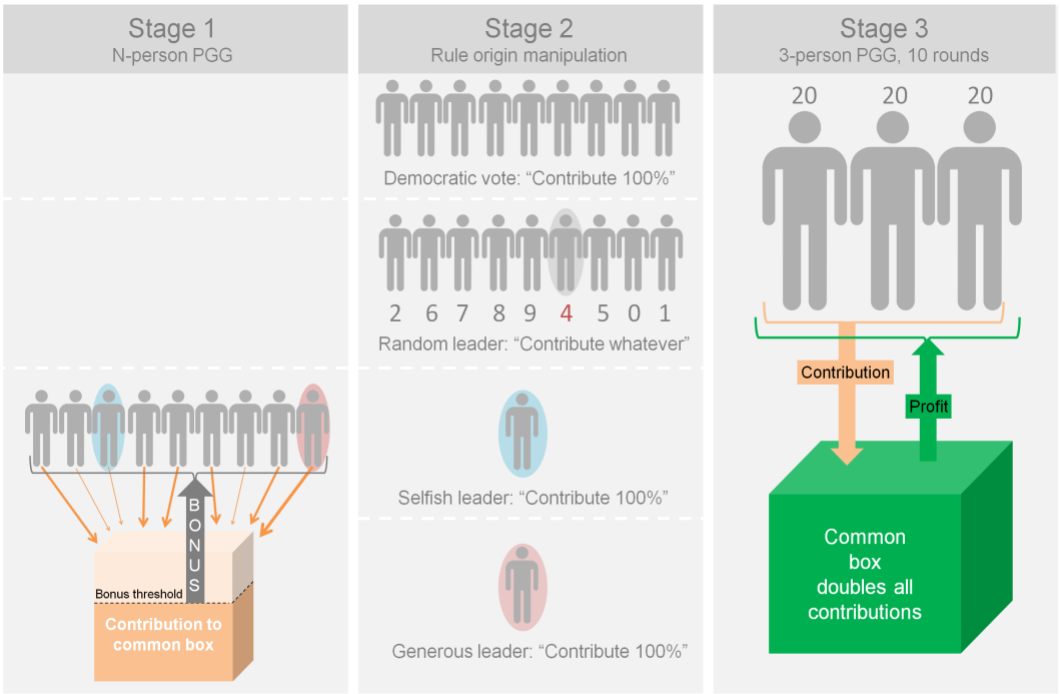


Figure 5.1. Session procedure for the different experimental conditions according to rule origin.

However, the way the rule was chosen varied according to the condition. In the democratic vote condition, the participants were instructed to vote for the rule they thought everyone in the session should follow and told that the rule would be decided by a majority vote. The voting procedure was private, with the rule being selected by clicking a button on the screen. After the voting was over, the rule was shown on participants' screens as well as re-read by the experimenter who was running the session to make sure that the rule was common knowledge, and to remove possible doubts participants might have had about everyone in the session not being shown the same rule. The rule presented to participants in the DV condition was always Rule 1, which implied maximal contributions to the common box.[16]

In the generous and selfish leader conditions, there was a pre-game in which participants could contribute to a session-wide common pot. If a certain threshold was reached, everyone would get the bonus in the end of the session, if not, there would be no bonus and the invested money to the common pot would be lost (participants were not aware whether the threshold was reached or not until the end of the session). Whatever participants chose to keep and not invest, they would get at the end of the session. After this game and the instructions about the game to follow, participants were told that either the person who had 'contributed the most' to the common pot (generous leader) or the person who had 'made the most money' (i.e. kept the most money units for themselves; selfish leader) would be the one choosing the rule.

We used the random leader condition as the baseline in this study. The 'random leader' was apparently chosen through a lottery process, in which each participant drew a random number out of a box and was asked to enter that number on their screen. Participants were told

[16] We pre-set this rule for all conditions in order to be able to compare the effect of rule origin on the contributions.

that the one with a number matching the pre-selected number of the leader would choose the rule. After putting in the number on the screen, all participants saw the same message – i.e. that they were not selected as the leader, and to click on the button to proceed and see which rule the leader had chosen. In reality, the rule was predetermined to Rule 5, 'Contribute whatever you want' (no rule).

After the rule was announced, participants proceeded to play ten rounds of the 3-person PGG as described above, with feedback about their profits and the total amount contributed to the common box after each round. Following the block of the PGG, participants were asked to provide their demographic information in a short questionnaire while their payments were being calculated. Figure 5.1. shows the different stages of the experiment for each of the conditions.

5.3. Results

5.3.1. Rule preferences

We ran a Chi-square goodness-of-fit test on the data we collected about rule preferences (N=66). The test showed a significant difference in the distribution of choices (as opposed to a uniform distribution; χ^2=55.364, df=4, p<.001). It is clear that this result comes from the majority of participants choosing Rule 1, which states that one should contribute their whole endowment to the common box in each round.

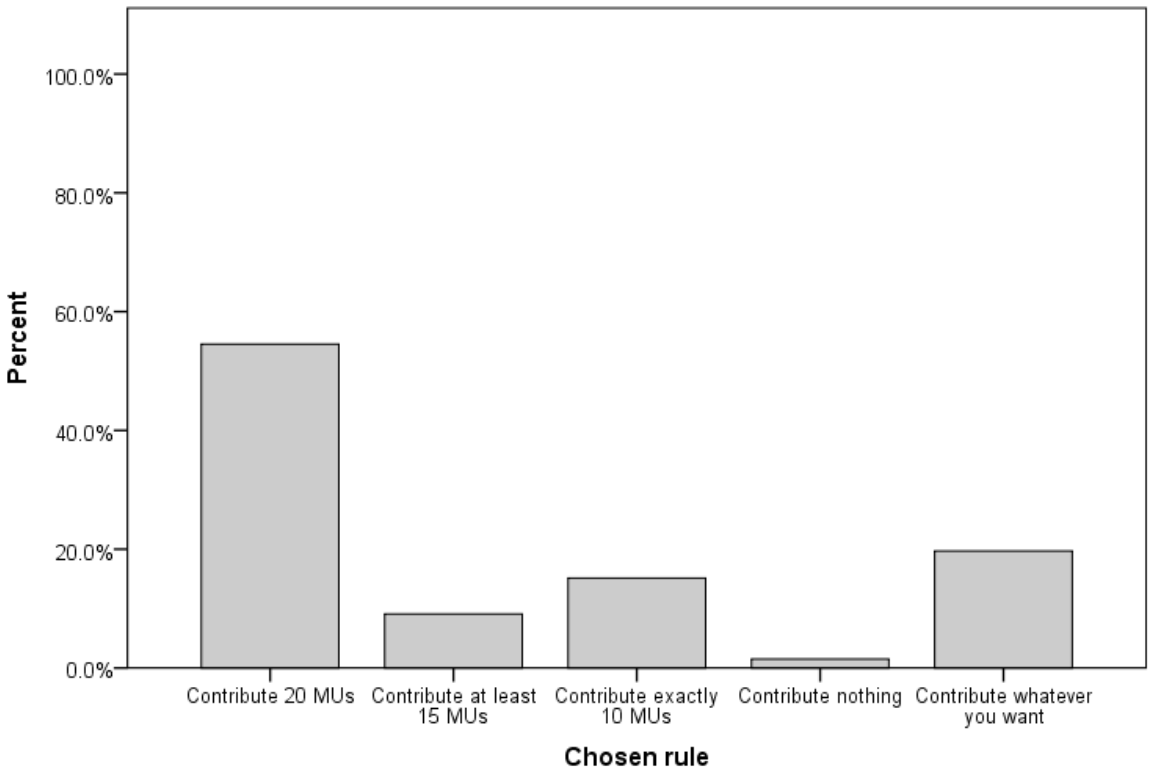


Figure 5.2. Percent of participants voting for each of the offered rules (data from Democratic Vote sessions).

5.3.2. Rule abidance and cooperation in the democratic vote condition

We first looked at initial rule abidance (in round 1), which was the main conceptual test of whether the origin of the rule had an influence on the participants' decisions. We compared the proportion of participants choosing to contribute 100% of their endowment in the democratic condition (as per the rule) with the proportion of participants doing the same in our 'baseline' condition (random leader who chose the 'give whatever you want' rule). Our analysis showed that the democratically chosen rule had an effect on maximal contributions in the first round (N=96, χ^2= 7.602, df=1, p=.007, Phi=0.2814): 52.26% participants chose to contribute their whole endowment to the common box in the democratic leader condition, whereas only 30.95% chose to abide by the rule in the baseline. This was also reflected in the initial contributions, which

were significantly higher in the democratic vote condition than in the baseline (Mann-Whitney *U*=824.000, *Z*=-2.426, *p*=.015, Cohen's d=.481).

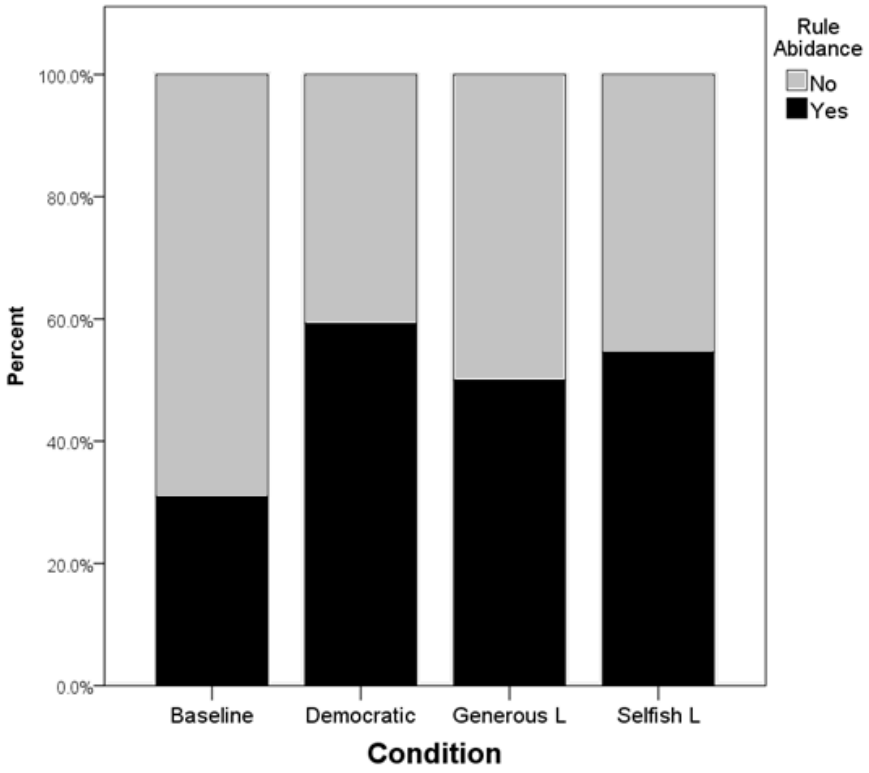


Figure 5.3. Initial rule abidance across the four conditions.

To inspect the influence of rule origin on rule abidance across rounds, we ran a generalized linear mixed-effects model (logit link) with participants nested into groups as the random effect; rule following as the dependent variable (coded as 0-didn't follow the rule, and 1-followed the rule); and condition (random leader as base), round (entered as a continuous predictor, round 1 as the base) and their interaction as fixed effects (see Table 5.1.). This, as well as all other mixed-effects analyses reported in this chapter, used the lme4 package in R (Bates et al., 2015).

There was a significant main effect of round, which reflected the decreasing likelihood of following the rule with each subsequent trial (*OR*=0.794, *SE*=0.053, *z*= -3.43, *p*<.001, 95% CIs [0.696, 0.906]). The effect of condition was only marginally significant, but pointed in the direction of more 100% contributions in the democratic condition.

Model	Rule abidance	Contributions
Fixed	Log-Odds (SE)	Estimate (SE)
Intercept (*base:random leader, R1*)	-1.88 (0.983)	12.505 (1.251)***
Condition (*democratic*)	2.195 (1.252)°	2.577 (1.696)
Round	-0.231 (0.067)***	-0.425 (0.091)***
Condition x Round	-0.077 (0.085)	-0.289 (0.121)*
Random		
Participant:group var(SD)	6.391 (2.528)	14.51 (3.809)
Group var(SD)	6.595 (2.568)	13.34 (3.653)
Observations, Participants, Groups	N=960, N=96, N=32	N=960, N=96, N=32
Model Fit	AIC= 726.4	AIC= 6173.2
	BIC= 755.6	BIC= 6207.3
	logLik=-357.2	logLik= -3079.6
	deviance= 714.4	deviance= 6159.2

Table 5.1. Mixed effects models of rule following (logit) and overall contributions across rounds in the democratic and baseline conditions. Significance levels: ° p < .10 * p < .05 ** p < .01 *** p < .001.

Finally, we were also interested in the overall cooperation levels – one of our predictions concerned the possible backfiring effect of the democratic rule if it was not followed. The model presented in Table 5.1. again shows a significant effect of round in the direction of lower contributions in later rounds, but no main effect of condition. However, there was a significant interaction between condition and round, such that contributions in the democratic vote condition seem to have been more adversely affected than those in the random leader condition.

5.3.3. Rule abidance and cooperation in the selfish and generous leader conditions

As with the democratic vote, we first analysed the data from the initial rounds for the selfish and generous rule-maker conditions, comparing them to the baseline. There was a significant difference between the selfish and random leader conditions (*N*=108, χ^2= 5.764, *df*=1, *p*=.019, Phi=0.231): whereas 30.95% participants in the baseline contributed the maximal

amount to the common box in the first round, 54.56% did so in the selfish leader condition. However, there was no significant difference between the mean contributions in the two conditions. Interestingly, there was no significant difference in rule-following between the baseline and the generous leader conditions (N=78, χ^2= 2.937, df=1, p=.107). Similarly, the difference between initial contributions in the two conditions was not significant (see Figure 5.4. for the mean contributions in each round across conditions).

A generalized linear mixed-effects model with rule following as the dependent variable, participants nested into groups as random effects, and condition (baseline, selfish leader, generous leader), round and the interaction of condition and round showed no significant main effect of condition or the interactions. The only significant fixed effect was that of round, which reflected the decreasing rule following in later rounds (OR=0.796, SE=0.053, z=-3.43, p<.001, 95% CIs [0.699, 0.907]). The same was true for the model predicting individual contributions to the common box, where again, only the round had a significant (negative) effect on the contributions (see Table 5.2.).

Model	Rule abidance	Contributions
Fixed	Log-Odds (SE)	Estimate (SE)
Intercept (*base:random leader, R1*)	-1.999 (0.856)*	12.079 (1.194)***
Condition (*generous leader*)	1.110 (1.198)	-0.185 (1.757)
Condition (*selfish leader*)	1.592 (1.048)	1.062 (1.527)
Round	-0.228 (0.067)***	-0.425 (0.091)***
Condition GL x Round	-0.012 (0.095)	0.040 (0.134)
Condition SL x Round	-0.050 (0.082)	-0.128 (0.116)
Random		
Participant:group var	6.590	18.46
Group var	4.612	10.51
Observations, Participants, Groups	N= 1440, N=144, N=48	N= 1440, N=144, N=48
Model Fit	AIC= 1078.2	AIC= 9264.7
	BIC= 1120.4	BIC= 9312.1
	logLik=-531.1	logLik= -4623.3
	deviance= 1062.2	deviance= 9246.7

Table 5.2. Mixed effects models of rule following (logit) and overall contributions across rounds in the selfish leader, generous leader and baseline conditions. Significance levels: * p < .05 ** p < .01 *** p < .001.

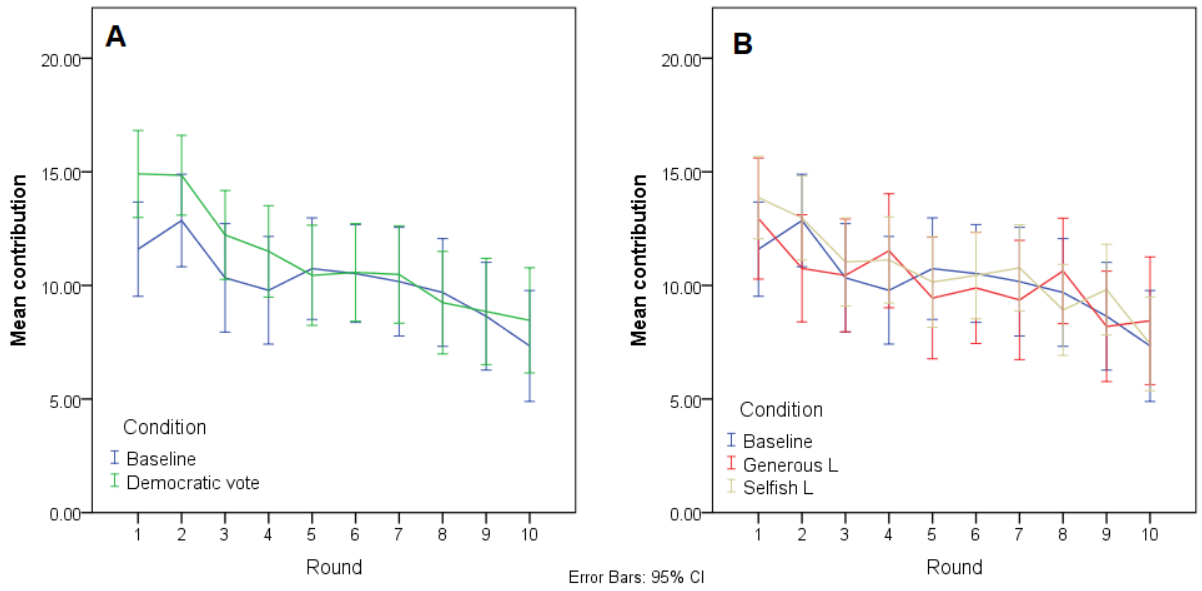


Figure 5.4. Mean contributions across rounds in the democratic vote (A) and generous and selfish leader conditions (B), compared to the baseline. Error bars represent 95% CIs.

5.4. Discussion

The aim of this study was to investigate the effect of the origin of a chosen rule on the willingness to abide by it. Our rationale was based partly on partner choice theory and the idea of the sceptical audience – we explored whether the same intuitions about the motivations of generous and selfish actors' prosocial choice (or in this case, declarations about which rules should be followed) translate into attitudes about said rules and affect rule abidance. We further investigated the effect of democratically chosen rules, which provide information about what one's partners (and audience) think one should do in the situation – how it affects the initial willingness to follow the rule and whether it has consequences on overall contributions during repeated rounds.

Our results show an initial willingness to follow the democratic rule – as well its negative effect on contributions in subsequent rounds. This is likely the case because participants observe that not everyone (in their group) abides by the rule and thus lower their contributions – even more than they would when no such rule has been stated and chosen by the group, because they might feel 'duped' by the audience choosing a costly rule they themselves do not follow (something akin to hypocrisy). We should state here that although the costly contribution rule was pre-selected to make comparisons between sessions in the same condition possible, a majority of participants did, in fact, vote for the maximal contribution, so the inference that they might've done so out of self-interest isn't completely unwarranted.

When it comes to the effect of leader intentions on rule abidance, we find an unexpected positive effect of the *selfish* leader on initial rule abidance, and no effect of the generous leader (which is the opposite to what we predicted). There are several post-hoc explanations we can offer to this end. For one, we note that the consequences of following the rule in our experiment

were transparent: there was no ambiguity or 'fudging' space which would make it seem like the intentions of the leader could affect one's earnings in the game or that the decision to choose the maximal contribution was otherwise underhanded. Future studies should thus implement more opacity in terms of the link between the rule and its effect on one's earnings in the game. For instance, the possibility of leaders collecting premiums on those who choose to follow the rule could be one such experimental manipulation which might bring the leader's intentions to the fore and influence the willingness to abide by the rule they chose. However, even this might be insufficient, if one considers the research on conflicts of interest and the 'perverse' effects of disclosure which show people often fail to discount biased advice as much as they should (Cain, Loewenstein, & Moore, 2011).

Another explanation has to do with framing effects – it is possible that qualifying the selfish leader as the participant 'who made the most profit' in the communal game (a choice we made in order to refrain from explicitly attributing selfishness to the leader and making experimenter demand effects more likely) was understood by our participants as an attribute of someone who knew how to play the game well and gain the most from it. In other words, instead of selfishness, they might've instead attributed competence in the given task and thus followed the rule based on this inference. This could also explain why the generous leader's suggested rule wasn't followed: their qualification of being prosocial might've been perceived as irrelevant (or even ill-suited) for making decisions in an economic game in which, presumably, most participants were motivated to earn as many points which would translate to money at the end of the session. Apart from the framing of how the leader was presented, the context of an 'economic experiment' might have thus played a part in which rule origin seemed relevant to take into account. It would be interesting to see how rule origins – especially with regard to rule-makers'

intentions – would interact with different, intentional frames imposed on the game itself, such as the 'Wall Street Game' and the 'Community Game' from Liberman et al.'s study (2004). We would venture to predict that such framing would affect *whose* rules participants would be more likely to follow.

Finally, it is also possible that participants did attribute selfish intentions to the leader, but that this lead them to think about the consequences of the rule more thoroughly as well as lead to more focus on the best strategy to maximize group welfare (which would then have translated into choosing to contribute maximally, at least in the first round).

It should also be noted that our sample consisted mostly of international students with heterogeneous cultural and ethnic backgrounds, which might have been a confound in terms of abiding by rules of different origins. Previous studies have shown that the type of rules participants prefer to follow can depend on their cultural context – for instance, Vollan, Landmann, Zhou, Hu and Herrmann-Pillath (2017) found that exogenously imposed rules (mimicking an 'authoritarian' context) were the most successful in increasing levels of cooperation in a sample of Chinese participants. It is possible that beyond the cultural influence on the contents of (pro)social rules, they also influence underlying inferences about the legitimacy and justifiability of rules when their origins are considered. This is another interesting avenue for future research: reputational and impression management considerations should be in-tune not only with the content of community-relevant social rules (like we've touched on in the previous chapter), but also with the community standards of what constitutes a legitimate rule as a more generalizable mechanism of 'rules about rules' which would make it easier for its members to gage which (new) rules they are expected to follow.

Chapter 6: Does assortment increase prosocial rule abidance?

6.1. Introduction

We have tested a number of hypotheses related to prosocial choice and rule abidance based on the assumption that a partner choice ecology is at the root of the mechanisms which drive these behaviours. Following this overarching rationale behind the thesis, we chose to address the effect of assortment (partner-matching) on prosocial rule abidance in this final chapter. We present the results of a simple experiment in which abiding by a rule was used to determine who one would interact with in the future – and contrast this with the instance where the rule had no such consequences. The rules used for the matching procedure were differed in that they were either 'cheap' to follow or costly – thus providing more certainty that the future partner would act prosocially, having been willing to pay a high cost to follow the rule. Our main hypothesis in this regard is that assortment will foster rule abidance – perhaps especially – in the case of high-cost rules because people will prefer to interact with those who are similarly labelled as 'rule followers' when following the rule provides sufficient evidence they will act prosocially with them, as well.[17]

Why do people abide by social rules, even at a cost to themselves? One reason can be an evolved norm-psychology, which presupposes an intrinsic preference for abiding by and enforcing the rules of one's community (e.g. Chudek & Henrich, 2011, see also Bicchieri, 2005). In this view, rule following is one consequence of being part of a community. By contrast, we show that rule following can also be a *means* to join a community, especially when it is

[17] Another reason (which we don't directly address in this study) would be a preference to think of oneself as a 'rule follower', i.e. the effect that being in this group might have on one's self-esteem, if the group is judged positively.

advantageous for one to do so. In other words, rule abidance need not reflect internalized norms or shared values, but can be used strategically in order to assort with those with whom it is fruitful to do so.

Partner choice ecology has been shown to foster prosociality in experimental settings (Barclay & Willer, 2007; Sylwester & Roberts, 2010) and through agent based simulation (e.g. Debove, André, & Baumard, 2015). We investigate a specific application of a partner choice ecology, characterized by assortative matching based on rule abidance. We predict that this ecology will foster rule abidance and thus prosocial choice on the assumption that (1) people will prefer to cooperate with those who are willing to confer benefits to others (including themselves), and (2) that they will be willing to invest in acting prosocially to gain access to valuable partners (Barclay, 2013). The assortative matching we implemented pairs those who choose to follow the rule with others who choose the same. Importantly, it also provides information about one's own willingness to incur high costs to act prosocially (by virtue of the matching), thus having the added benefit of serving as an indicator of one's prosociality, particularly in the high-cost rule where uncertainty about the absolute contribution is much lower than in the low-cost rule condition.

We had strong predictions about the effect of the matching procedure on the willingness to abide by the rule, however, when it came to the cost of the rule we initially considered several plausible outcomes they might have on rule abidance (and especially cooperation). On the one hand, the threshold of what is considered 'prosocial' (given the rule) might boost cooperation in the costly rule condition, leading to more people contributing high amounts to their partners to be matched with those who did the same. On the other hand, it might also have an adverse effect of

leading those who would've otherwise contributed less than 8 (but not zero), to shift towards the lower end of the scale, including zero (following a 'go big or go home' strategy).

With regard to the cheap rule, we considered that providing the justification (in form of the small rule) for contributing little could have an effect even without the rule being announced – something akin to 'hiding behind the small cake' (Ockenfels & Werner, 2012) where most contributions would cluster around the minimal prescribed amount. Secondly, it could also serve to nudge those who would otherwise contribute nothing to send at least some of their initial endowment in order to abide by the rule, decreasing zero contributions when assortment is implemented. Thirdly, we also considered the possibility that the cheap rule would be disregarded due to its lack of value in informing about the follower's prosociality, i.e. that participants would contribute as if didn't exist (meaning, no clustering around the rule).

We try to disentangle some of these options by considering the graphs and the analyses of the experimental data below.

6.2. Method

6.2.1. Participants

Participants were recruited through the SONA system at the Central European University, with the condition of proficiency in English. A total of 108 participants signed up and completed the experiment (54 female, 35 male, 3 other; *M* age=26.64[18]), 52 (28 in the assortment-first and 24 in the assortment-second sessions) in the 30% rule condition and 56 the 80% condition (26 in the assortment-first and 30 in the assortment-second sessions). The experiment was approved by the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

[18] Please note that demographic data is missing for 16 participants due to technical error.

6.2.2. Experimental design

The experiment consisted of forty rounds of the Continuous Prisoner's dilemma (PD) game. The game is played in pairs: each participant gets an endowment of ten points, any amount of which they can send to their partner. The amount sent to the partner is doubled by the experimenter in the transfer, so that the payoff from a round for one participant equals the points they keep in the round plus double the points they are sent by their partner (see Figure 6.1. for an example round).

There were two distinct phases (blocks) of the experiment: random matching blocks, consisting of twenty rounds of the game where one's contribution in the previous round had no influence on who one's partner in the subsequent round would be (they were paired randomly); and assortment blocks, equally consisting of twenty rounds, but in this case the contributions in each round affected who the participant would be paired with in the next round. Specifically, rule abidance was used as a matching tool, so that participants who followed the rule in the current round were paired with others who did the same. By the same token, those who didn't follow the rule in the current round were matched with others who didn't follow the rule in the same round on the next turn. (If there was an uneven number of rule followers/non-followers, one participant's contribution from each pool was duplicated to calculate earnings for the unmatched participants.) The blocks were counterbalanced across sessions, and represent the within-subjects factor of the experimental design.

The between-subjects factor was the rule, i.e. the amount of money units one needed to send to the partner in a given round to be in the 'rule-followers' group. In one condition, we set this rule to 30% of one's endowment (i.e. a contribution of 3 or more money units), whereas in the other condition it was set to 80% of the endowment (a contribution of 8 or more money units).

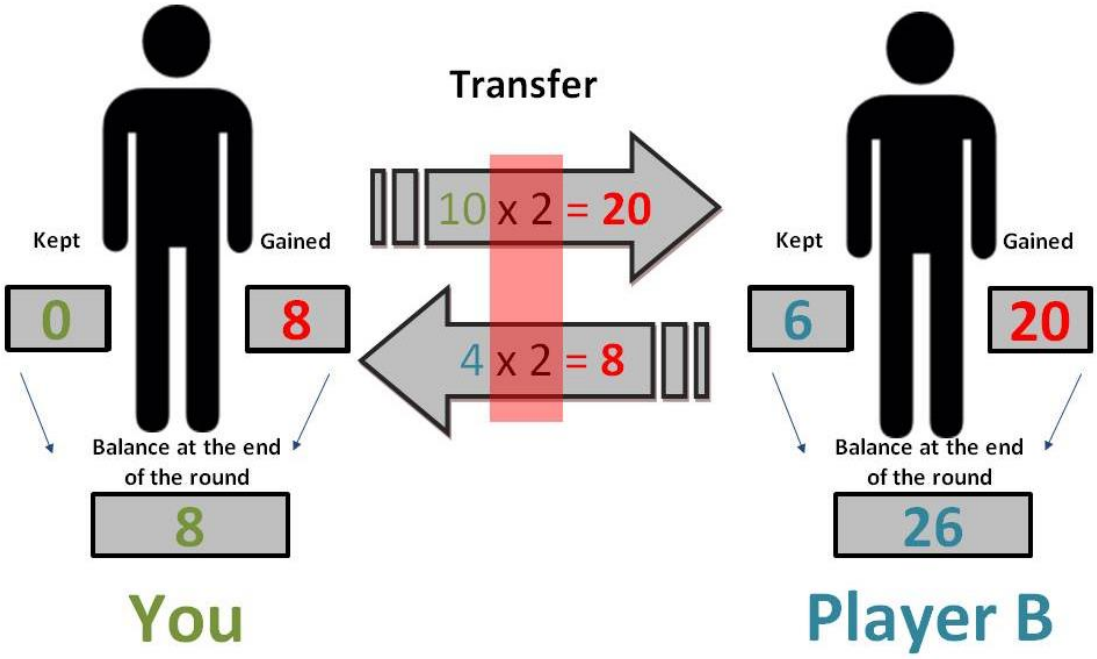


Figure 6.1. An example of a round shown to participants, explaining the payoff calculations.

6.2.3. Procedure

The experiment was held in the computer labs of the Central European University, using the Z-Tree software (Fischbacher, 2007). Each session lasted 30-40 minutes. Participants were paid a show-up fee of 500 HUF, which was added to the profits they made during the game. Upon arrival, participants were seated at their work stations, which were separated by white dividers on three sides (left, right and front of the computer) to provide anonymity. Each participant was given an identifying number for the length of the experiment, which they later used to receive their payments from the game.

The instructions for the experiment were presented on the computer screen as well as read out loud by the experimenter, with a prompt for questions at the end of the introductory section. Participants were then asked to solve a series of comprehension questions about imaginary rounds of the game to ensure their understanding of the game. After each comprehension check, they received feedback with the correct answers and an explanation of

how these correct answers were calculated. At the end of the comprehension checks, we further prompted participants to ask questions if they had any uncertainties about the game.

After this practice phase, participants were informed about the rule (according to the condition they were in), which was framed as a suggestion from the experimenters, and for which no further justification was given. Depending on which phase came first (assortment or no assortment), they were also asked to answer two questions about the matching procedure (whether two players with different contributions could be matched in the next round, according to the rule). They answered these questions either before the first phase, if it was the assortment phase; or in-between the first and second phase, after they received further instructions about the matching procedure. However, note that the rule itself was always presented at the end of the instructions section, i.e. before participants started playing the game.

At the end of the experiment, participants were asked to answer a short questionnaire collecting their demographic information while their payoffs were being calculated.

6.3. Results

6.3.1. The effect of assortment on initial abidance and contributions

We first analysed the data from the initial rounds of the assortment and no assortment blocks separately, with the rule (30 vs. 80%) as the row and rule abidance as the column (followed vs. didn't follow the rule) of a 2x2 contingency table. There was a significant effect of rule costliness on the ratio of abidance in the no assortment blocks (χ^2=7.155, df =1, p = .007, Phi=0.257): the participants in the 30% rule condition were more likely to follow it than those in the 80% condition even when there was no influence of matching. However, this difference was not significant in the assortment blocks (χ^2=2.965, df =1, p= .085), suggesting that the participants in the 80% rule condition found the cost of following the rule worth it when it meant

they would be matched with others who had also been willing to send 8 or more MUs in the initial round.

To further test for the effect of assortment, we ran Repeated-Samples McNemar's tests on the paired rule-abidance decisions between the no-assortment and assortment blocks in both conditions. Both tests were significant, reflecting the increased number of participants abiding by the rule in the assortment blocks, in both the 30% Rule condition (χ^2=50.019, df=1, p<.001) and the 80% Rule condition (χ^2=54.018, df=1, p<.001).

Though rule abidance was higher in the 30% condition, we also wanted investigate the effect of rules on initial contributions. Specifically, we wanted to see whether no assortment coupled with a high-cost rule would drive cooperation down and result in lower mean contributions due to more participants choosing the zero option. Figure 6.2. shows the average contribution amounts for each condition and block combination (across rounds). Looking at the no-assortment lines for each rule, it appears there is no difference in contributions between the two rules in the first round: this is also the result we get by comparing them with a Mann-Whitney U tests (N=108, Mann-Whitney U=1299.5, p=.322). However, we do find a significant difference in contributions in the assortment blocks (N=108, Mann-Whitney U=1072.500, p=0.017, d = 0.466) – participants in the Rule 80% condition have higher contributions, presumably because of the higher proportion of those choosing to follow the rule.

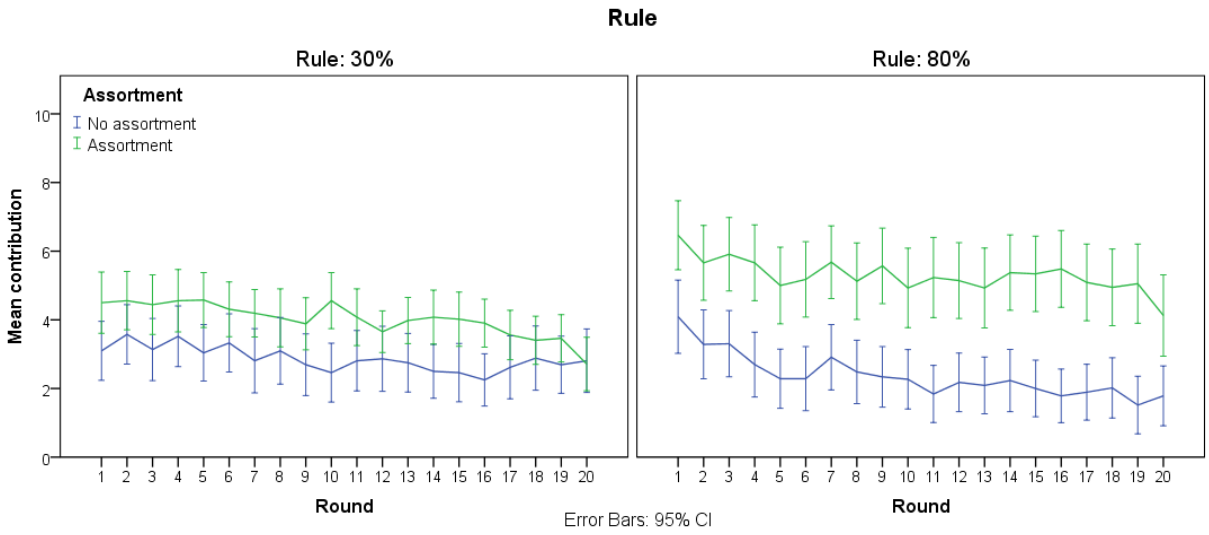


Figure 6.2. Mean contributions across assortment and no assortment rounds, in Rule 30% and Rule 80% conditions. Error bars represent 95% confidence intervals.

These results seem to point against the first stated option, i.e. that the costly rule coupled with no assortment would lead to a decrease in cooperation in comparison to the cheap-to-follow rule – at least initially. Still, it is possible that this difference comes from a distribution in which more people were clustered on the two tail-ends of the distribution, i.e. that there were more participants who either gave 8 and more MUs *and* more participants who gave zero. Figure 6.3. shows the percentage (per condition) of initial contribution choices in the assortment and no assortment blocks.

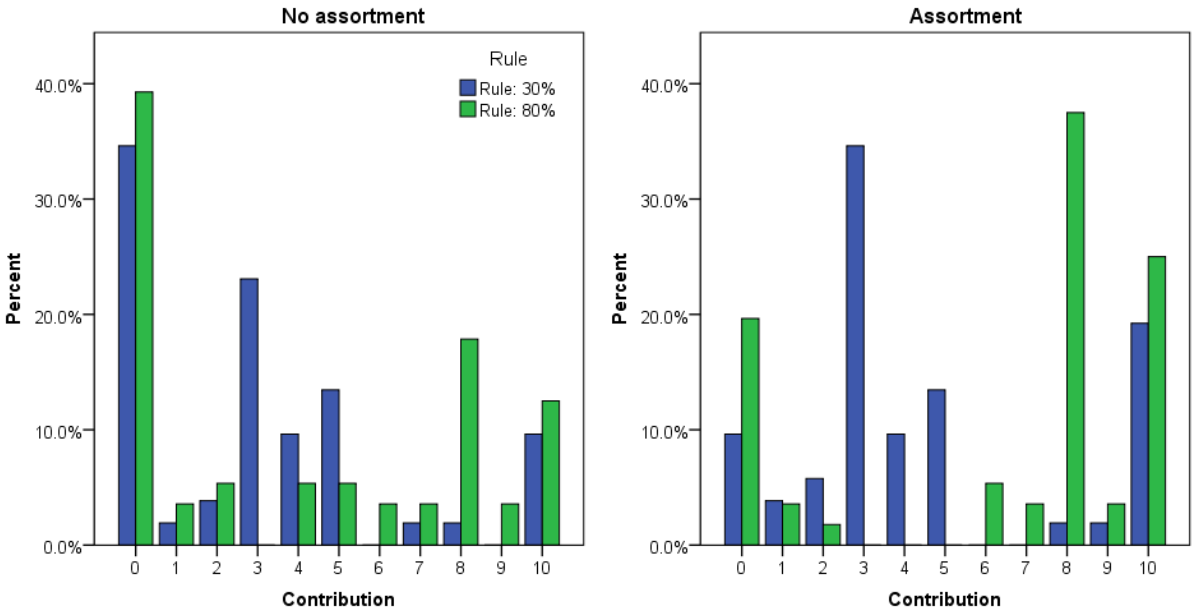


Figure 6.3. Percentage of participants choosing each contribution amount in Round 1, per condition.

Looking at the no assortment block, the frequency of choosing the zero contribution does not appear to be significant (and in fact, it is not when comparing the proportions of zero vs. non-zero contributions by rule in the no assortment initial round – χ^2= 0.252, p=.616). The same is true of the assortment block, where the difference is more pronounced, but still not significant (χ^2= 2.148, p=.143). The driver of the (no) difference seem to be the contributions in the 30% rule condition, especially in the assortment condition where the most frequent transfer was 3 MUs, whereas the most common contributions in the 80% rule condition were 8 MUs and 10 MUs, respectively. Looking at the graphs, this seems to be a consequence of those previously contributing zero contributing 3 MUs in the assortment block. However, it is also interesting to note that even in the 30% condition, a non-negligible number of participants still chose to contribute more than the minimal amount, as well as send the maximum amount to their partners.

6.3.2. Comparison of rule abidance across assortment and no-assortment rounds

The average percent of rule-following (across rounds) in the rule 30% condition was 52.02% in the no assortment blocks, and 80.87% in the assortment blocks. In the 80% condition, the average percent of rule-following in the no assortment block was 13.40%, which increased to 55.98% in the assortment condition.

In order to compare the effect of the rule on abidance across rounds, we looked at the two rules separately using mixed-effects models with condition (no assortment as base) and round (introduced as a continuous predictor) as the fixed factors, their interaction, and participants as random effects to account for the repeated measures aspect of the design. The analyses were done in R, using the lme4 R package (Bates et al., 2015).

In the model of 30% rule data, all main effects were significant: there was a higher likelihood of following the 30% rule in the assortment condition (*OR*=4.28, *SE*=1.14, *z*=5.47, *p*<.001, 95% CI [2.54, 7.21]), an effect of round (*OR*=0.955, *SE*=0.0098, *z*=-4.48, *p*<.001, 95% CI [0.936, 0.974]) which showed a small decline of rule-following as the game progressed. The counterbalancing was also significant (*OR*=.321, *SE*=.121, *z*=-2.79, *p*=.006, 95% CI [.0828, .646]), showing that rule following decreased when the assortment condition came first in the session. The interaction between assortment condition and counterbalance condition proved not to be significant.

We implemented the same analysis on the data from the 80% condition. The estimates again showed that there was a significantly higher likelihood of following the 80% rule (*OR*=127.74, *SE*=48.4, *z*= 12.8, *p*<.001, 95% CI [60.90, 269.00]), with an effect of round showing that rule following overall decreased across rounds (*OR*=0.940, *SE*=0.0118, *z*=-4.94, *p*<.001, 95% CI [0.917, 0.963]). Unlike previously, the counterbalance order was not significant,

but there was a significant interaction between counterbalance and assortment (*OR*=0.264, *SE*=0.171, *z*= -2.05, *p*=0.040, 95% CIs [0.0743, 0.941]), reflecting that presentation order influenced contributions only in the assortment condition.

Model	Rule: 30%	Rule: 80%
Fixed	Log-Odds (SE)	Log-Odds (SE)
Intercept (*base: no assortment, assortment second, round 1*)	1.921 (0.370)***	-2.052 (0.637)**
Condition (*assortment*)	1.454 (0.266)***	4.851 (0.379)***
Counterbalance (*assortment first*)	-1.464(0.524)**	-0.946 (0.921)
Round	-0.046 (0.010)***	-0.062 (0.013)***
Condition x Counterbalance	0.654 (0.469)	-1.330 (0.648)*
Random		
ID var	2.625	9.208
Observations, Participants	N=2080, N=52	N=2240, N=56
Model Fit	AIC= 1908.4	AIC= 1439.0
	BIC= 1942.3	BIC= 1473.3
	logLik= -948.2	logLik= -713.5
	deviance= 1896.4	deviance= 1427.0

Table 6.1. Generalized linear mixed model for rule following fit by maximum likelihood (Laplace Approximation), family: logit. Significance levels: * p < .05 ** p < .01 *** p < .001.

6.3.3. Comparison of contributions across assortment and no-assortment rounds

The mean of averaged contributions across rounds for participants in the no assortment block of the rule 30% condition was *M*=2.869, and *M*=3.964 in the assortment rounds. In the rule 80%, the means were, respectively, *M*=2.541 and *M*=5.295 (refer back to Figure 6.2. for a visual representation of mean contributions across rounds).

To further investigate the effect of the rule on contributions, we ran generalized mixed-effects models on the two rules. The model for the assortment data again included condition, counterbalancing, their interaction and round as fixed predictors and participants as a random effect (see Table 6.2.). The analysis showed that the condition had a significant effect on contributions, i.e. they were higher in the assortment condition in comparison to the base. Round was also a significant predictor, with the contributions expectedly decreasing as participants

progressed through the game. Finally, there was no interaction between counterbalancing and condition, but a general increase in contributions for both conditions when the assortment condition was presented first.

Results in the 80% rule condition also mirrored rule following results presented above. The condition had a significant effect on the contributions alongside the round, both in the predicted direction. There was no significant effect of main counterbalancing, but an interaction between condition and counterbalancing indicating that order influenced the contributions for the assortment condition only, specifically resulting in lower contributions when it was presented first compared to second.

Model	Rule: 30%	Rule: 80%
Fixed	Estimate (SE)	Estimate (SE)
Intercept (*base:no asssortment, assortment second, round 1*)	4.883 (0.420)***	4.355 (0.540)***
Condition (*assortment*)	1.173 (0.566)***	3.730 (0.242)***
Counterbalance (*assortment first*)	-1.445(.583)*	-0.976 (0.751)
Round	-0.058 (0.007)***	-0.073 (0.001)***
Condition x Counterbalance	-0.266 (0.358)	-1.686 (0.426)***
Random		
ID var	3.870	7.073
Observations, Participants	N=2080, N=52	N=2240, N=56
Model Fit	AIC= 9072.3 BIC= 91111.8 logLik= -4529.2 deviance= 9058.3	AIC= 10718.4 BIC= 10758.4 logLik= -5352.2 deviance= 10704.4

Table 6.2. Linear mixed model for contributions fit by maximum likelihood (t-tests use Satterthwaite's method). Significance levels: * p < .05 ** p < .01 *** p < .001.

6.4. Discussion

The main aim of the study we presented in this chapter was to gauge the effect of assortment on the willingness to abide by prosocial rules, even when said rules are costly. We hypothesized that the structure of institutions where sorting is allowed will push more players to contribute higher amounts than they would normally, so that they can belong to the group that

fosters the better social equilibrium. In other words, that assortment should lead to more rule-following than simply stating a suggested rule which has no consequences on neither one's payoffs nor future partnerships. This should especially be seen as the case when the rule is high enough to indicate the partner's – and in turn one's own – willingness to invest in mutually beneficial interactions.

This main prediction was reflected in our results – participants were indeed willing to pay a high cost in order to 'enter the pool' of rule-followers when it meant their partner would be selected from the same pool, whereas the same was not true when no matching procedure was implemented. Our results showed that, overall, the patterns of rule-abidance and contributions in the two conditions (cheap and costly rules) differed in the assortment and no-assortment blocks. While the costly rule was initially less likely to be followed than the cheap rule when there was no partner-matching, this difference disappeared in the first round of the assortment blocks. Furthermore, there were no differences in the mean contributions between the 30% and the 80% conditions in the first round of the no assortment blocks, whereas the initial contributions in the assortment block differed as a result of the increased number of rule-followers in the Rule 80% condition. Additional analyses of the first rounds also showed there was no significant 'push' towards zero contributions in the costly rule condition, which we considered as a plausible scenario. In fact, mean contributions in the 80% Rule condition were either higher (assortment) or no different (no assortment) than those in the cheap rule condition.

We also noted the non-negligible amount of participants who contributed more than the rule, especially in the assortment blocks where zero-contributions were decreased. This could be related to strategic decisions to access cooperative partners (especially in the costly rule condition) by those who'd given zero in the no-assortment blocks. It could also conceivably be a

combination of both the participants' social preferences and their self-image concerns, the latter affecting their choice in a similar way it would on a biological market, i.e. to be better than the minimum – especially when the minimum is not informative – in order to attract valuable partners.

We witness a small, but significant decrease in cooperation and rule abidance across rounds in both condition. There is also an order effect of the blocks, such that counterbalancing negatively influenced abidance in the 30% Rule condition (there was more rule-following overall when assortment followed the no-assortment block). In the 80% Rule condition, the significant interaction between assortment and counterbalancing indicated that the order effect decreased contributions only in the assortment blocks, i.e. that participants were slightly less likely to follow the rule in the assortment block if it came first. This is interesting inasmuch as it shows that lower contributions in the no-assortment blocks likely provide further incentives for participants to invest in rule-abidance in the subsequent assortment rounds (whereas no such 'additional' incentive exists when the session starts with the assortment block).

Chapter 7: Conclusions

After six chapters of studies examining the various factors which influence cooperativeness, attributions of prosociality and rule abidance, what can we conclude or predict about the Napoletan tradition of suspended coffee (apart from the fact that the residents of Italy love the beverage as much to consider it essential enough to provide as gifts for those unable to pay for it themselves)?

We might say that providing this gift is more likely when someone knows about it – especially if there is a large number of patrons one expects to meet again – perhaps next morning – to witness the action. Even more importantly, if some of those patrons are potential business partners who value benevolence and charitable behaviour, or potential acquaintances one would like to impress. While making the gesture, one should also beware to do it in a way as to not draw attention to oneself (for instance, by adding it to one's tab while paying for one's own coffee in hush-hush voices without making a spectacle of it), counting on the staff to gossip about it or for interested parties to pay close attention to the action. One might also consider where the tradition originates from while deciding whether to engage in it or not – and take into account the number of times they've seen others do it in turn. Finally, the price of the coffee might also be an important factor in the decision – and interact with the abovementioned number of observers and the importance the potential coffee-giver places on their reputation with said observers.

What this little imagined field experiment should make clear is that the relationship between impression management and signalling on the one hand and prosociality, cooperation and rule-following on the other is multi-faceted and complex, driven by often conflicting

motivations and sensitive to the context in which the interactions are embedded, which in and of itself is characterized by a variety of features that need to be taken into account if prosocial impression management is to be adaptive. In this thesis, we provide a diverse set of evidence for the 'signatures' of these cognitive mechanisms which do just that: enable people to successfully use information about the relevant factors and maximize their benefits from future interactions in impression management situations.

The focus put on observers' individual differences in the theoretical chapter reflects the main goal of this thesis, which has been to provide an account of adaptive strategies for making prosocial impressions on a biological market (Noë & Hammerstein, 1995, Barclay, 2016) populated by strategically vigilant observers (Heintz, Karabegovic, & Molnar, 2016). To this end, we looked at both the influence of audience characteristics on the willingness to signal prosociality, and the attributions of prosociality inferred from contexts with different strategic value for the actor. The conclusions of our empirical studies can thus be divided into three main themes, which are peppered and often presented together throughout the chapters, but which we here discuss separately: (1) impression management in prosocial and rule-abidance contexts; (2) strategic vigilance, attributions of prosociality and partner choice; and (3) the influence of rule features and affordances on rule-abidance and cooperation. In the final section, we discuss the limitations of the studies and possible directions for future research.

7.1. Impression management in prosocial and rule-abidance contexts

Chapters 2 and 3 directly examined the influence of audience features on prosocial choice, while Chapter 4 indirectly touched upon impression management and self-enhancement effects on the stated willingness to follow social rules. The study of overt and covert changes to contributions in view of new information about observation and potentially beneficial future

interactions from Chapter 2 is the most direct test of actors' intuitions about what we've referred to as the '*Catch-22*' of managing prosocial impressions – the difference in changed contributions between the private and public knowledge conditions implies people are aware of how strategic prosociality will be perceived by observers, and that they can adjust their strategies accordingly.

To illustrate the point, take these two comments taken from the exit questionnaire after the experimental sessions. One participant from the public knowledge condition said the following: "*I put 0 into the common box, so I thought I would not have been chosen with that amount for sure. I changed it to 8. I did not want to change it to 10, because that way in the eyes of the green group I would be just somebody who is obviously wanting to be chosen to put 0 to the common box afterwards*." On the other hand, a participant from the private knowledge condition remarked that they "*wanted to show a better side of me so I'd be picked again. Not particularly for the money (but that was part of it) but mostly to be seen better than the other and to be chosen again...to be loved?*" These statements are interesting for several reasons – for one, because they provide further evidence of meta-cognition in this context and show participants had similar hypotheses about how a change in their prosocial contributions could be interpreted (and thus didn't go for the maximal contribution in the public knowledge condition – which is an interesting and very psychologically-minded strategy). What the other statement so insightfully summarizes, however, is that a mechanism of impression management is likely to operate through proximal mechanisms as well as pure conscious strategy: how our participant put it, the need to be better than the other, to be chosen or loved. Of course, money is also a factor; however it is relegated to a secondary consideration, after social competition, partner choice and social approval.

Chapter 3 tested a more straightforward prediction: that audience 'quality' would be taken into account when deciding whether or not to split an initial endowment fairly or keep it to oneself. Again, the main prediction we made – that observation by a 'high-quality' audience, i.e. one with whom the first player can later make up the cost of the prosocial signal with would be more likely to increase the proportion of signalling prosociality than no observation or observation by a 'low-quality' audience – was borne out of the data. Interestingly, we also found a similar effect in the passive audience condition, namely, when participants were made aware someone would see their decision, but weren't told about future interactions with the observer, they acted similarly as when being observed with a high-quality audience they expected to interact with.

It is possible that experimenter demand had a part to play in this instance, i.e. that our participants had reasoned that the observer would not have been mentioned unless there was some future contingency which included them they weren't yet aware of (which was true). However, even in this case, the fact that the imagined audience was a *relevant* one bears mentioning. Behaving as if observers are relevant until given evidence to the contrary would likely have constituted an adaptive impression management mechanism evolutionarily speaking, especially in partner choice ecologies where most observers were likely to be at least somewhat important for one's reputation, if not through direct benefits from future interactions with themselves, then through transmitting reputational information through gossip.

Finally, the results from the study presented in Chapter 4, which investigates the willingness to coordinate on and self-report (allegedly) unethical behaviours in view of their acceptability to the local community and the framing in which they are presented, also point to the influence of audience values in impression management. Specifically, we show that not all

seemingly unethical choices are made equal – some are more likely to be reported in hypothetical scenarios even in the first person, while some are not. The chapter also provides insights about the potential uses of coordination games in the study of social rules and discusses the benefits of combining methods (such as incentivized games and self-reports, in this case) for the study of attitudes towards different rules and the underlying motivations for abiding by the same.

7.2. Strategic vigilance, attributions of prosociality and partner choice

Corresponding to the analyses of the influence of observer features on the willingness to manage prosocial impressions, Chapters 2 and 3 also looked at the other side of the coin, namely, the dependence of inferences about actors' prosociality and their desirability as potential partners on the context of surrounding the prosocial display. Data gathered from observers in the study about private and public strategic changes showed that unchanged contributions (those less likely to be strategic) were preferred over changed contributions – the latter were only preferred when the outside options were patently uncooperative or (relatedly) when the difference between the changed an unchanged contributions was very high.

The second study in Chapter 2, which investigated attributions of prosocial traits and liking in hypothetical scenarios differing with regard to audience relevance and cost of the prosocial action, showed that the perceived relevance of the audience observing the prosocial action played a part in deciding whether it was strategically motivated, and reflected on subsequent attributions of prosocial traits, liking and predictions of future prosocial choices (which didn't afford reputational benefits). However, we did not find the effect in one of the three hypothetical scenarios, which differed from the others in that the contribution was that of time and effort as opposed to money, and that the prosocial action was directed at a friend, rather

than a public (office) good or a charity. This null-result raises interesting questions about how the identity of and relationship to the receiver of a certain prosocial action (if such exists as an individual) influence the perception of the action as well as how it's categorized, as well as raising the issue of the commensurability of different types of prosociality. Regarding the latter, we find it likely that certain types of prosocial actions are more likely to fall prey to discounting than others, especially when viewed as 'cheap' (e.g. clicking on a like button, sharing a post on Facebook or even contributing 5 euro to support a cause vs. investing time and effort to volunteer).

The results from observer groups in the audience quality experiment (Chapter 3) did not exhibit the same scepticism we found in the abovementioned studies. Though we did find an influence of players' previous generosity on observers' trust decisions, in that cooperative players were trusted more than uncooperative ones, the trust shown to cooperative players was not qualified by condition, as we'd predicted (that those who were aware of the strategic incentives to be generous would be met with more scepticism). Partner choice preferences generally followed our predictions in terms of preferring the players who displayed generosity without being observed, though we also found an interesting preference for the cooperative players from the high-quality condition over those from the low-quality condition (following the logic of strategic considerations, those who gave more without expecting a return should be judged as more prosocial, perhaps even than those in the no-audience condition). One possible explanation is that observers predicted that those who expected to be receive more in the Trust game would've also been more likely to decide on a higher back-transfer.

Similarly, in our study of the influence of leader strategic intentions on rule-abidance and cooperation, which was based on the same rationale of sceptical audiences, we find the opposite

of the predicted effect – the rules decided by selfish leaders were more likely to be followed than those made by generous leaders. We consider the various plausible explanations of this result at length in the discussion section of Chapter 5.

Taken together, while the results summarized above do provide some evidence for strategic vigilance in audiences, the effect seems to be more subtle than we'd expected. It is likely that at least some of the null-results are due to methodological and experimental design features (e.g. in Chapter 5) and that these underlying hypotheses would benefit from conceptual replications which take our initial efforts into account.

7.3. The influence of rule features and affordances on rule-abidance and cooperation

The last set of studies in this thesis explored how rule origins, the costliness of the rule and its use in partner-matching affect rule abidance. We already touched on the unpredicted results that the intentions of leaders had on rule abidance above – the other part of the study in Chapter 5 was related to democratically chosen rules, which we considered as a 'special' origin case due to their inherent informativeness about audience values and expectations. Our results showed that rules chosen by majority vote do positively influence initial rule abidance, however, we also find that they also seem to lead to marginally quicker decrease in overall cooperation (represented by the amount of contributions to the public good) across rounds.

In the experiment described in Chapter 6, we find that implementing assortment can increase rule-following even when the rules call for costly transfers. We also find a non-negligible number of participants are willing to contribute *more* than the rule dictates, which – we speculate – could be evidence of competitive altruism and impression management adapted

to a biological market where simply matching the minimal threshold of what is considered acceptable often isn't enough to secure the best partnerships. On the other hand, 'cheap' rules seem to have the effect of inciting those who would otherwise give zero to match the minimal threshold in order to gain access to the pool of 'cooperators'. If we were to speculate and generalize these results to policy-making, we believe this distinction would be interesting to take into account for 'nudgers' when tailoring measures to increase rule-abidance. Specifically, our results can be taken to imply that rule costs should depend on the target population one is trying to reach: those who disregard the rule completely, or those who show some initial willingness to pay the cost of abidance.

7.4. Future directions

While discussing the results from the studies included in this thesis, we've already mentioned (or hinted at) possible extensions or improvements to the experimental protocols which could build on the current results and further distinguish between the plausible explanations we've provided for the more puzzling findings. We summarize them here, outlining what we perceive to be the potentially most interesting avenues for future research.

To expand on the above section in which we've already touched upon this subject in terms of rule-abidance, apart from implementing experimental designs which make intentions more important for participants' payoffs (i.e. by increasing the opacity of a rule's consequences and the self-interests of the leader), further studies might pit one rule against another – both in terms of costliness and origins – to see which are preferentially followed when 'in competition'. For example, when a rule decided on by the majority differs from a rule suggested by a leader (with one or another trait relevant to the context). This, in tandem with frames imposed on the games as previously mentioned, could provide relevant insights about the underlying motivations for

rule-following and their hierarchy, similar to studies which pitted descriptive against injunctive norms.

Another surprising result which could be elaborated on is the effect of 'passive' audiences, i.e. observers whose potential future importance for an actor isn't known. One caveat of studies employing vignettes and hypothetical scenarios is that the information on which the decision is hypothesized to rely on has to be provided (more or less) explicitly. However, uncertainty about the social actors one comes into contact with abounds in daily life. We are often ill-equipped to parse such factors as audience quality or probability of future interactions at a glance, and what's more, appearances can often be deceiving and lead to misjudgments. As elaborated on in Chapter 1, error management theory (Haselton & Galperin, 2012) would likely guide one to make conservative predictions about the value of potential partners, erring on the side of caution. Are the audiences we imagine thus relevant by default? Do we manage impressions based on this assumption until given concrete evidence to the contrary (similar to waning effects of eye-cues, such as discussed in Sparks & Barclay, 2013)? What constitutes as evidence to the contrary, and how much evidence is enough to discount observation? These questions have mostly been tackled indirectly so far, and would benefit from investing more concentrated research efforts in.

Finally, we'd like to underscore two issues which are particularly relevant for impression management in the current *Zeitgeist*: the exponential increase of opportunities for self-presentation (to heterogeneous audiences, no less, which is a third aspect that deserves attention in its own right) via online social networks and the resulting degradation of signal value when signals are cheap to produce. The latter is also related to the abovementioned commensurability of prosocial actions (and consequently signals), but more to the point, it refers to a shift in the

perceived market (and audience) size, as well as in what constitutes a reliable or adequate cue of one's prosociality.

Studies such as the 'humble-bragging' one by Sezer, Gino and Norton (2018) have pointed to the apparent missteps in impression management in online contexts, among others. Investigating the strategies which form in answer to this conspicuous strategic dimension of self-disclosures on social networking sites – and the strategies used to compete for reputations – is bound to be a fruitful avenue for researchers interested in signalling prosociality. Not only, but certainly also because of the wealth of available real-world data that is waiting to be explored.

References

Alekseev, A., Charness, G., & Gneezy, U. (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behaviour & Organization, 134*, 48-59.

Alexander Jr, C. N., & Rudd, J. (1981). Situated identities and response variables. In J.T. Tedeschi (Ed.), *Impression management theory and social psychological research*. NY: Academic Press.

Alexander Jr, C. N., & Weil, H. G. (1969). Players, persons, and purposes: Situational meaning and the prisoner's dilemma game. *Sociometry, 32*(2), 121-144.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of personality and social psychology, 49*(6), 1621-1630.

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. *The self in social judgment, 1*, 85-106.

Alloway, T., Runac, R., Quershi, M., & Kemp, G. (2014). Is Facebook linked to selfishness? Investigating the relationships among social media use, empathy, and narcissism. *Social Networking, 3*(3), 150-158.

Andrade, E. B., & Ho, T. H. (2009). Gaming emotions in social interactions. *Journal of Consumer Research, 36*(4), 539-552.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal, 100*(401), 464-477.

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica, 77*(5), 1607-1636.

Andreoni, J., & Petrie, R. (2004). Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of public Economics, 88*(7-8), 1605-1623.

Andrews, P. W. (2001). The psychology of social chess and the evolution of attribution mechanisms: Explaining the fundamental attribution error. *Evolution and Human Behaviour, 22*(1), 11-29.

Aquilino, W. S. (1997). From adolescent to young adult: A prospective study of parent-child relations during the transition to adulthood. *Journal of Marriage and the Family 59*(3), 670-686.

Aquilino, W. S., Wright, D. L., & Supple, A. J. (2000). Response effects due to bystander presence in CASI and paper-and-pencil surveys of drug use and alcohol use. *Substance Use & Misuse, 35*(6-8), 845-867.

Ariely, D. (2012). *The (honest) truth about dishonesty.* New York, NY: HarperCollins.

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544-55.

Arkin, R. M. (1981). Self-presentation styles. In J.T. Tedeschi (Ed.), *Impression management theory and social psychological research*. New York, NY: Academic Press.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*, *70*(9), 1-70.

Ashton, M. C., Paunonen, S. V., Helmes, E., & Jackson, D. N. (1998). Kin altruism, reciprocal altruism, and the Big Five personality factors. *Evolution and Human Behaviour*, *19*(4), 243-255.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of economic perspectives*, *29*(3), 3-30.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological bulletin*, *140*(6), 1556-81.

Barasch, A., Berman, J. Z., & Small, D. A. (2016). When payment undermines the pitch: On the persuasiveness of pure motives in fund-raising. *Psychological Science*, *27*(10), 1388-1397.

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behaviour*, *34*(3), 164-175.

Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current opinion in psychology*, *7*, 33-38.

Barclay, P., & Reeve, H. K. (2012). The varying relationship between helping and individual quality. *Behavioural Ecology*, *23*(4), 693-698.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, *274*(1610), 749-753.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.). (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press, USA.

Bateson, M., Callow, L., Holmes, J. R., Roche, M. L. R., & Nettle, D. (2013). Do images of 'watching eyes' induce behaviour that is more pro-social or more normative? A field experiment on littering. *PloS one*, *8*(12), e82055.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology letters*, *2*(3), 412-414.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioural and Brain Sciences*, *36*(1), 59-78.

Baumeister, R. E., & Tice, D. M. (1986). Four selves, two motives, and a substitute process self-regulation model. In R. E. Baumeister (Ed.), *Public self and private self.* New York, NY: Springer.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The quarterly journal of economics*, *117*(3), 871-915.

Berger, C., & Rodkin, P. C. (2012). Group influences on individual aggression and prosociality: Early adolescents who change peer affiliations. *Social Development*, *21*(2), 396-413.

Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The Braggart's dilemma: On the social rewards and penalties of advertising prosocial behaviour. *Journal of Marketing Research*, *52*(1), 90-104.

Bermúdez, J.L. (2020). Frame It Again: New Tools for Rational Decision-Making. Cambridge: Cambridge University Press.

Bersoff, D. M. (1999). Why good people sometimes do bad things: Motivated reasoning and unethical behaviour. *Personality and social psychology bulletin*, *25*(1), 28-39.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioural Decision Making*, *22*(2), 191-208.

Bird, R. B., & Power, E. A. (2015). Prosocial signalling and cooperation among Martu hunters. *Evolution and Human Behaviour*, *36*(5), 389-397.

Bird, R. B., Smith, E., & Bird, D. W. (2001). The hunting handicap: costly signalling in human foraging strategies. *Behavioural Ecology and Sociobiology*, *50*(1), 9-19.

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, *41*(4), 540-558.

Blascovich, J., Mendes, W. B., Hunter, S. B., & Salomon, K. (1999). Social" facilitation" as challenge and threat. *Journal of personality and social psychology*, *77*(1), 68-77.

Bó, P. D. (2005). Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American economic review*, *95*(5), 1591-1604.

Boen, F., Vanbeselaere, N., & Feys, J. (2002). Behavioural consequences of fluctuating group success: An Internet study of soccer-team fans. *The Journal of social psychology*, *142*(6), 769-781.

Bond, C. F. (1982). Social facilitation: A self-presentational view. *Journal of Personality and Social Psychology, 42*(6), 1042-1050.

Bond, Jr, C. F., & Titus, L. J. (1983). Social facilitation: a meta-analysis of 241 studies. *Psychological bulletin, 94*(2), 265-292.

Bond, Jr, C. F., Atoum, A. O., & VanLeeuwen, M. D. (1996). Social impairment of complex learning in the wake of public embarrassment. *Basic and applied social psychology, 18*(1), 31-44.

Bond, R. (2005). Group size and conformity. *Group processes & intergroup relations, 8*(4), 331-354.

Boone, J. L. (1998). The evolution of magnanimity. *Human Nature, 9*(1), 1-21.

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist 6*(4), 308-311.

Bosson, J. K., Taylor, J. N., & Prewitt-Freilino, J. L. (2006). Gender role violations and identity misclassification: The roles of audience and actor variables. *Sex Roles, 55*(1-2), 13-24.

Bourdage, J. S., Roulin, N., & Levashina, J. (2017). Impression management and faking in job interviews. *Frontiers in psychology, 8*(1294), 1-4.

Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality?. *Proceedings of the Royal Society B: Biological Sciences, 285*(1875), 20180116.

Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social cognition, 4*(4), 353-376.

Burnham, T. C., & Hare, B. (2007). Engineering human cooperation. *Human nature, 18*(2), 88-108.

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: an evolutionary perspective on human mating. *Psychological review, 100*(2), 204-232.

Cain, D. M., Loewenstein, G., & Moore, D. A. (2011). When sunlight fails to disinfect: Understanding the perverse effects of disclosing conflicts of interest. *Journal of Consumer Research, 37*(5), 836-857.

Cappelen, C., & Dahlberg, S. (2018). The Law of Jante and generalized trust. *Acta Sociologica, 61*(4), 419-440.

Casale, S., & Banchi, V. (2020). Narcissism and problematic social media use: A systematic literature review. *Addictive Behaviours Reports, 11*, 100252.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics, 117*(3), 817-869.

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in cognitive sciences, 15*(5), 218-226.

Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of personality and social psychology, 34*(3), 366-375.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social influence, 1*(1), 3-15.

Coricelli, G., Rusconi, E., & Villeval, M. C. (2014). Tax evasion and emotions: An empirical test of re-integrative shaming theory. *Journal of Economic Psychology, 40*, 49-61.

Corral-Verdugo, V., Frias-Armenta, M., Pérez-Urias, F., Orduña-Cabrera, V., & Espinoza-Gallego, N. (2002). Residential water consumption, motivation for conserving water and the continuing tragedy of the commons. *Environmental management, 30*(4), 527-535.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 163–228. Oxford, UK: Oxford University Press.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308-315.

Cronk, L. (2007). The influence of cultural framing on play in the trust game: A Maasai example. *Evolution and Human Behaviour, 28*(5), 352-358.

Cronk, L., & Wasielewski, H. (2008). An unfamiliar social norm rapidly produces framing effects in an economic game. *Journal of Evolutionary Psychology, 6*(4), 283-308.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology, 40*, 61-149.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behaviour. *Personnel Psychology, 47*(4), 817-829.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behaviour and human decision Processes, 100*(2), 193-201.

Davis, D. W., & Silver, B. D. (2003). Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science, 47*(1), 33-45.

De Freitas, J., DeScioli, P., Thomas, K. A., & Pinker, S. (2019). Maimonides' ladder: States of mutual knowledge and the perception of charitability. *Journal of Experimental Psychology: General, 148*(1), 158-173.

Dear, K., Dutton, K., & Fox, E. (2019). Do 'watching eyes' influence antisocial behaviour? A systematic review & meta-analysis. *Evolution and Human Behaviour, 40*(3), 269-280.

Debove, S., André, J. B., & Baumard, N. (2015). Partner choice creates fairness in humans. *Proceedings of the Royal Society B: Biological Sciences, 282*(1808), 20150392.

Delton, A. W., & Robertson, T. E. (2012). The social cognition of social foraging: Partner selection by underlying valuation. *Evolution and human behaviour, 33*(6), 715-725.

DeScioli, P., & Kurzban, R. (2009). The alliance hypothesis for human friendship. *PloS one, 4*(6), e5802.

DeScioli, P., & Kurzban, R. (2011). The company you keep: Friendship decisions from a functional perspective. In Krueger, J. I. (Ed.), *Social Judgment and Decision Making.* New York, NY: Psychology Press.

Diener, E. (1977). Deindividuation: Causes and consequences. *Social Behaviour & Personality: an international journal, 5*(1), 143-156.

Dubreuil, B., & Grégoire, J. F. (2013). Are moral norms distinct from social norms? A critical assessment of Jon Elster and Cristina Bicchieri. *Theory and Decision, 75*(1), 137-152.

Duffy, J., & Ochs, J. (2009). Cooperative behaviour and the frequency of social interaction. *Games and Economic Behaviour, 66*(2), 785-812.

Dufwenberg, M., & Muren, A. (2006). Generosity, anonymity, gender. *Journal of Economic Behaviour & Organization, 61*(1), 42-49.

Ehlebracht, D., Stavrova, O., Fetchenhauer, D., & Farrelly, D. (2018). The synergistic effect of prosociality and physical attractiveness on mate desirability. *British Journal of Psychology, 109*(3), 517-537.

Eisenbruch, A. B., & Roney, J. R. (2017). The Skillful and the stingy: Partner choice decisions and fairness intuitions suggest human adaptation for a biological market of cooperators. *Evolutionary Psychological Science, 3*(4), 364-378.

Elster, J. (2007). *Explaining social behaviour: More nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.

Ensminger, J. (2000). Experimental economics in the bush: How institutions matter. In Menard, C. (Ed.), Institutions and organizations. London: Edward Elgar.

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behaviour: a field experiment. *Evolution and Human Behaviour*, *32*(3), 172-178.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioural, and biomedical sciences. *Behaviour research methods*, *39*(2), 175-191.

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

Fehr, E., & Schneider, F. (2010). Eyes are on us, but nobody cares: are eye cues relevant for strong reciprocity?. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1686), 1315-1323.

Fein, S., & Hilton, J. L. (1994). Judging others in the shadow of suspicion. *Motivation and Emotion*, *18*(2), 167-198.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological science*, *25*(3), 656-664.

Feltovich, N., Harbaugh, R., & To, T. (2002). Too cool for school? Signalling and countersignalling. *RAND Journal of Economics*, 630-649.

Ferrari, J. R., & Díaz-Morales, J. F. (2007). Perceptions of self-concept and self-presentation by procrastinators: Further evidence. The Spanish journal of psychology, 10(1), 91-96.

Finch, J. F., & Cialdini, R. B. (1989). Another indirect tactic of (self-) image management: Boosting. *Personality and Social Psychology Bulletin*, *15*(2), 222-232.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, *10*(2), 171-178.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., ... & Kainbacher, M. (2011). The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological bulletin*, *137*(4), 517-37.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. Journal of consumer research, 20(2), 303-315.

Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, *99*(4), 689- 723.

Fiske, S. T., & Berdahl, J. (2007). Social power. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles*. New York: Guilford Press.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences, 11*(2), 77-83.

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

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and individual differences, 7*(3), 385-400.

Gallier, C. (2020). Democracy and compliance in public goods games. *European Economic Review, 121*, 103346.

Galperin, A., & Haselton, M. G. (2012). Error management and the evolution of cognitive bias. In J. P. Forgas, K. Fiedler, & C. Sedikedes (Eds.), *Social thinking and interpersonal behaviour.* New York: Psychology Press.

Gambetta, D. (2011). *Codes of the underworld: How criminals communicate*. Princeton, NJ: Princeton University Press.

Gambetta, D., & Przepiorka, W. (2014). Natural and strategic generosity as signals of trustworthiness. *PloS one, 9*(5), e97533.

Gambetta, D., & Székely, Á. (2014). Signs and (counter) signals of trustworthiness. *Journal of Economic Behaviour & Organization, 106*, 281-297.

Gelfand, M. J., & Harrington, J. R. (2015). The motivational force of descriptive norms: For whom and when are descriptive norms most predictive of behaviour?*Journal of Cross-Cultural Psychology, 46*(10), 1273-1278.

Ghodsee, K. (2018). *Why women have better sex under socialism: and other arguments for economic independence.* New York, NY: Nation Books.

Gino, F., Norton, M. I., & Weber, R. A. (2016). Motivated Bayesians: Feeling moral while acting egoistically. Journal of Economic Perspectives, 30(3), 189-212.

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York, NY: Doubleday.

Grafen, A. (1990). Biological signals as handicaps. *Journal of theoretical biology, 144*(4), 517-546.

Guinote, A., Cotzia, I., Sandhu, S., & Siwa, P. (2015). Social status modulates prosocial behaviour and egalitarianism in preschool children and adults. *Proceedings of the National Academy of Sciences, 112*(3), 731-736.

Haley, K. J., & Fessler, D. M. (2005). Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human behaviour, 26*(3), 245-256.

Hamilton, A. F. D. C., & Lind, F. (2016). Audience effects: what can they tell us about social neuroscience, theory of mind and autism?. *Culture and Brain, 4*(2), 159-177.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of personality and social psychology, 78*(1), 81-91.

Haselton, M., & Galperin, A. (2012). Error management and the evolution of cognitive bias. *Soc. Think. Interpers. Behav. 45*, 63.

Hawkes, K., O'Connell, J. F., & Jones, N. G. B. (2014). More lessons from the Hadza about men's work. *Human Nature, 25*(4), 596-619.

Heintz, C. (2006). Web search engines and distributed assessment systems. *Pragmatics & Cognition, 14*(2), 387-409.

Heintz, C., Celse, J., Giardini, F., & Max, S. (2015). Facing expectations: Those that we prefer to fulfil and those that we disregard. *Judgment & Decision Making, 10*(6), 442–455.

Heintz, C., Karabegovic, M., & Molnar, A. (2016). The co-evolution of honesty and strategic vigilance. *Frontiers in psychology, 7*,1503.

Heldt, T. (2005). Sustainable Nature Tourism and the Nature of Tourists' Cooperative Behaviour: Recreation Conflicts, Conditional Cooperation and the Public Good Problem (Doctoral dissertation).

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... & Henrich, N. S. (2005). "Economic man" in cross-cultural perspective: Behavioural experiments in 15 small-scale societies. *Behavioural and brain sciences, 28*(6), 795-815.

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

Hilbig, B. E., & Zettler, I. (2009). Pillars of cooperation: Honesty–Humility, social value orientations, and economic behaviour. *Journal of Research in Personality, 43*(3), 516-519.

Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts towards cooperation as a matter of dispositional honesty–humility. *European Journal of Personality, 26*(3), 245-254.

Hoff, K., & Pandey, P. (2006). Discrimination, social identity, and durable inequalities. *American Economic Review, 96*(2), 206-211.

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behaviour in dictator games. *The American economic review, 86*(3), 653-660.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences, 112*(6), 1727-1732.

Jellison, J. M. (1981). Reconsidering the attitude concept: A behaviouristic self-presentation formulation. In J.T. Tedeschi (Ed.), *Impression management theory and social psychological research.* NY: Academic Press.

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. In J. Suls (Ed.), *Psychological perspectives on the self.* Hillsdale, NJ: Erlbaum.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences, 113*(31), 8658-8663.

Kafashan, S., Sparks, A., Griskevicius, V., & Barclay, P. (2014). Prosocial behaviour and social status. In J.T. Cheng, J.L. Tracy, C. Anderson (Eds.), *The psychology of social status*. New York, NY: Springer.

Kindelan, K. (2017, June 21). 167 drivers pay it forward in McDonald's drive-thru. ABC News. Retrieved from: https://abcnews.go.com/Lifestyle/167-drivers-pay-forward-mcdonalds-drive/story?id=48180780

Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social cognition, 27*(4), 623-634.

Krendl, A., Gainsburg, I., & Ambady, N. (2012). The effects of stereotypes and observer pressure on athletic performance. *Journal of Sport and Exercise Psychology, 34*(1), 3-15.

Krueger, J. (1998). Enhancement bias in descriptions of self and others. *Personality and Social Psychology Bulletin, 24*(5), 505-516.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity, 47*(4), 2025-2047.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?. *Journal of the European Economic Association, 11*(3), 495-524.

Kuran, T. (1987). Preference falsification, policy continuity and collective conservatism. *The Economic Journal, 97*(387), 642-665.

Kurzban, R. (2011). *Why everyone (else) is a hypocrite*. Princeton, NJ: Princeton University Press.

Lamba, S., & Mace, R. (2013). The evolution of fairness: Explaining variation in bargaining behaviour. *Proceedings of the Royal Society B: Biological Sciences, 280*(1750), 20122028.

Latane, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of personality and social psychology, 10*(3), 215- 221.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin, 107*(1), 34-47.

Leary, M. R., Raimi, K. T., Jongman-Sereno, K. P., & Diebels, K. J. (2015). Distinguishing intrapsychic from interpersonal motives in psychological theory and research. *Perspectives on Psychological Science, 10*(4), 497-517.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate behavioural research, 39*(2), 329-358.

Lesorogol, C. K. (2007). Bringing norms in: The role of context in experimental dictator games. *Current anthropology, 48*(6), 920-926.

Li, N. P., van Vugt, M., & Colarelli, S. M. (2018). The evolutionary mismatch hypothesis: Implications for psychological science. *Current Directions in Psychological Science, 27*(1), 38-44.

Libby, R., & Rennekamp, K. (2012). Self-serving attribution bias, overconfidence, and the issuance of management forecasts. *Journal of Accounting Research, 50*(1), 197-231.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and social psychology bulletin, 30*(9), 1175-1185.

Lin-Healy, F., & Small, D. A. (2012). Cheapened altruism: Discounting personally affected prosocial actors. *Organizational Behaviour and Human Decision Processes, 117*(2), 269-274.

Lin-Healy, F., & Small, D. A. (2013). Nice guys finish last and guys in last are nice: The clash between doing well and doing good. *Social Psychological and Personality Science, 4*(6), 692-698.

Malkis, F. S., Kalle, R. J., & Tedeschi, J. T. (1982). Attitudinal politics in the forced compliance situation. *The Journal of Social Psychology, 117*(1), 79-91.

Marder, B., Joinson, A., Shankar, A., & Thirlaway, K. (2016). Strength matters: Self-presentation to the strongest audience rather than lowest common denominator when faced with multiple audiences in social network sites. *Computers in Human Behaviour, 61*, 56-62.

Martin, R., & Randal, J. (2008). How is donation behaviour affected by the donations of others?. *Journal of Economic Behaviour & Organization, 67*(1), 228-238.

Matsugasaki, K., Tsukamoto, W., & Ohtsubo, Y. (2015). Two failed replications of the watching eyes effect. *Letters on Evolutionary Behavioural Science, 6*(2), 17-20.

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.

Mehta, J., Starmer, C., & Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review, 84*(3), 658-673.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and personality psychology compass, 4*(5), 344-357.

Mifune, N., Hashimoto, H., & Yamagishi, T. (2010). Altruism toward in-group members as a reputation mechanism. *Evolution and Human Behaviour, 31*(2), 109-117.

Milgram, S., Bickman, L., & Berkowitz, L. (1969). Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology, 13*(2), 79-82.

Milinski, M., Semmann, D., Bakker, T. C., & Krambeck, H. J. (2001). Cooperation through indirect reciprocity: image scoring or standing strategy?. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 268*(1484), 2495-2501.

Miton, H., & Mercier, H. (2015). Cognitive obstacles to pro-vaccination beliefs. *Trends in Cognitive Sciences, 19*(11), 633-636.

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass, 11*(8), e12331.

Nettle, D., & Saxe, R. (2020). Preferences for redistribution are sensitive to perceived luck, social homogeneity, war and scarcity. *Cognition, 198*, 104234.

Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological science, 25*(3), 648-655.

Noë, R., & Hammerstein, P. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioural ecology and sociobiology, 35*(1), 1-11.

Noë, R., & Hammerstein, P. (1995). Biological markets. *Trends in Ecology & Evolution, 10*(8), 336-339.

Northover, S. B., Pedersen, W. C., Cohen, A. B., & Andrews, P. W. (2017). Artificial surveillance cues do not increase generosity: Two meta-analyses. *Evolution and Human Behaviour, 38*(1), 144-153.

O'Grady, C.*, Kliesch, C.*, Smith, K., & Scott-Phillips, T. C. (2015). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution & Human Behaviour, 36*(4), 313-322.

Ockenfels, A., & Werner, P. (2012). 'Hiding behind a small cake' in a newspaper dictator game. *Journal of Economic Behaviour & Organization, 82*(1), 82-85.

Oda, R., Niwa, Y., Honma, A., & Hiraishi, K. (2011). An eye-like painting enhances the expectation of a good reputation. *Evolution and Human Behaviour, 32*(3), 166-171.

Ostrom, E., & Nagendra, H. (2006). Insights on linking forests, trees, and people from the air, on the ground, and in the laboratory. *Proceedings of the national Academy of sciences, 103*(51), 19224-19231.

Peeters, B. (2004). Thou shalt not be a tall poppy": Describing an Australian communicative (and behavioural) norm. *Intercultural Pragmatics, 1*(1), 71-92.

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature communications, 5*, 4939.

Pianigiani, G. (2014, December 24). In Naples, Gift of Coffee to Strangers Never Seen. The New York Times. Retrieved from: https://www.nytimes.com/2014/12/25/world/europe/naples-suspended-coffee.html

Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. *PloS one, 9*(2), e88534.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of personality and social psychology, 75*(3), 811-832.

Pleasant, A., & Barclay, P. (2018). Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychological science, 29*(6), 868-876.

Raihani, N. J., & Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society open science, 3*(11), 160510.

Raihani, N. J., & Smith, S. (2015). Competitive helping in online giving. *Current Biology, 25*(9), 1183-1186.

Read, D. (2005). Monetary incentives, what are they good for?. Journal of Economic Methodology, 12(2), 265-276.

Rege, M., & Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of public Economics, 88*(7-8), 1625-1644.

Richardson, K. D., & Cialdini, R. B. (1981). Basking in reflected glory. In J.T. Tedeschi (Ed.), *Impression management theory and social psychological research*. New York, NY: Academic Press.

Richter, N., Over, H., & Dunham, Y. (2016). The effects of minimal group membership on young preschoolers' social preferences, estimates of similarity, and behavioural attribution. *Collabra 2*(1), 1-8.

Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology, 30*(3), 358-367.

Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. Proceedings of the Royal Society of London. *Series B: Biological Sciences, 265*(1394), 427-431.

Robertson, T. E., Sznycer, D., Delton, A. W., Tooby, J., & Cosmides, L. (2018). The true trigger of shame: Social devaluation is sufficient, wrongdoing is unnecessary. *Evolution and Human Behaviour, 39*(5), 566-573.

Ronson, J. (2015). *So You've Been Publicly Shamed*. New York, NY: Riverhead Books.

Rotella, A. (2020). Who cooperates and why? Investigations of the roles of individual differences and reputation in cooperative behaviours (Doctoral dissertation). Available from University of Guelph (http://hdl.handle.net/10214/17893).

Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual review of psychology, 43*(1), 133-168.

Schneider, D. J. (1981). Tactical self-presentations: Toward a broader conception. In J.T. Tedeschi (Ed.), *Impression management theory and social psychological research.* New York, NY: Academic Press.

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3*(2), 102-116.

Sedikides, C., Herbst, K. C., Hardin, D. P., & Dardis, G. J. (2002). Accountability as a deterrent to self-enhancement: The search for mechanisms. *Journal of personality and social psychology, 83*(3), 592-605.

Seta, C. E., & Seta, J. J. (1995). When audience presence is enjoyable: The influence of audience awareness of prior success on performance and task interest. *Basic and Applied Social Psychology, 16*(1-2), 95-108.

Sezer, O., Gino, F., & Norton, M. I. (2018). Humblebragging: A distinct—and ineffective—self-presentation strategy. *Journal of Personality and Social Psychology, 114*(1), 52-74.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science, 24*(2), 125-130.

Sherman, R. A., Nave, C. S., & Funder, D. C. (2010). Situational similarity and personality predict behavioural consistency. *Journal of personality and social psychology, 99*(2), 330-343.

Smith, J. R., Louis, W. R., Terry, D. J., Greenaway, K. H., Clarke, M. R., & Cheng, X. (2012). Congruent or conflicted? The impact of injunctive and descriptive norms on environmental intentions. *Journal of Environmental Psychology, 32*(4), 353-361.

Soetevent, A. R. (2005). Anonymity in giving in a natural context—a field experiment in 30 churches. *Journal of public Economics, 89*(11-12), 2301-2323.

Sowden, S., Koletsi, S., Lymberopoulos, E., Militaru, E., Catmur, C., & Bird, G. (2018). Quantifying compliance and acceptance through public and private social conformity. *Consciousness and cognition, 65*, 359-367.

Sparks, A., & Barclay, P. (2013). Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behaviour, 34*(5), 317-322.

Sparks, E., Schinkel, M. G., & Moore, C. (2017). Affiliation affects generosity in young children: The roles of minimal group membership and shared interests. *Journal of experimental child psychology, 159*, 242-262.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language, 27*(5), 495-518.

Steele, C. M. (1975). Name-calling and compliance. *Journal of Personality and Social Psychology, 31*(2), 361-369.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology, 69*(5), 797-811.

Steinmetz, J., Sezer, O., & Sedikides, C. (2017). Impression mismanagement: People as inept self-presenters. *Social and Personality Psychology Compass, 11*(6), e12321.

Sundie, J. M., Kenrick, D. T., Griskevicius, V., Tybur, J. M., Vohs, K. D., & Beal, D. J. (2011). Peacocks, Porsches, and Thorstein Veblen: Conspicuous consumption as a sexual signalling system. *Journal of personality and social psychology, 100*(4), 664-680.

Swakman, V., Molleman, L., Ule, A., & Egas, M. (2016). Reputation-based cooperation: empirical evidence for behavioural strategies. *Evolution and Human Behaviour, 37*(3), 230-235.

Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology letters, 6*(5), 659-662.

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences, 113*(10), 2625-2630.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology, 1*(2), 149-178.

Tal-Or, N. (2010). Bragging in the right context: Impressions formed of self-promoters who create a context for their boasts. *Social Influence, 5*(1), 23-39.

Tedeschi, J. T., & Riess, M. (1981). Identities, the phenomenal self, and laboratory research. In J.T. Tedeschi (Ed.), *Impression management theory and social psychological research*. NY: Academic Press.

Tedeschi, J. T., Schlenker, B. R., & Bonoma, T. V. (1971). Cognitive dissonance: Private ratiocination or public spectacle?. *American Psychologist, 26*(8), 685-695.

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. Journal of personality and social psychology, 116(3), 396-415.

Tennie, C., Frith, U., & Frith, C. D. (2010). Reputation management in the age of the world-wide web. *Trends in cognitive sciences, 14*(11), 482-488.

Tetlock, P. E., & Manstead, A. S. (1985). Impression management versus intrapsychic explanations in social psychology: A useful dichotomy?. *Psychological review, 92*(1), 59-77.

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behaviour: A theoretical framework and meta-analysis. *Psychological Bulletin, 146*(1), 30-90.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press, USA.

Tracer, D. (2003). Selfishness and fairness in economic and evolutionary perspective: An experimental economic study in Papua New Guinea. *Current Anthropology, 44*(3), 432-438.

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *The American journal of psychology, 9*(4), 507-533.

Van Lange, P. A., Bekkers, R., Schuyt, T. N., & Vugt, M. V. (2007). From games to giving: Social value orientation predicts donations to noble causes. *Basic and applied social psychology, 29*(4), 375-384.

Vollan, B., Landmann, A., Zhou, Y., Hu, B., & Herrmann-Pillath, C. (2017). Cooperation and authoritarian values: An experimental study in China. *European Economic Review, 93*, 90-105.

Von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioural and Brain Sciences, 34*(1), 1-56.

Warneken, F., Sebastián-Enesco, C., Benjamin, N. E., & Pieloch, K. A. (2019). Pay to play: Children's emerging ability to use acts of generosity for selfish ends. *Journal of experimental child psychology, 188*, 104675.

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science, 288*(5467), 850-852.

Williams, K. E., Sng, O., & Neuberg, S. L. (2016). Ecology-driven stereotypes override race stereotypes. *Proceedings of the National Academy of Sciences, 113*(2), 310-315.

Wincenciak, J., Fincher, C. L., Fisher, C. I., Hahn, A. C., Jones, B. C., & DeBruine, L. M. (2015). Mate choice, mate preference, and biological markets: the relationship between partner choice and health preference is modulated by women's own attractiveness. *Evolution and Human Behaviour, 36*(4), 274-278.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1251-1263.

Wortman, C. B., Costanzo, P. R., & Witt, T. R. (1973). Effect of anticipated performance on the attributions of causality to self and others. *Journal of Personality and Social Psychology, 27*(3), 372-381.

Wu, J., Balliet, D., & Van Lange, P. A. (2016). Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific reports, 6*, 23919.

Wu, J., Balliet, D., & Van Lange, P. A. (2016). Reputation management: Why and how gossip enhances generosity. *Evolution and Human Behaviour, 37*(3), 193-201.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly, 63*(2), 116-132.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and emotion, 18*(2), 129-166.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., ... & Simunovic, D. (2013). Is behavioural pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behaviour and Human Decision Processes, 120*(2), 260-271.

Zahavi, A., & Zahavi, A. (1997). *The handicap principle: a missing piece of Darwin's puzzle*. New York, NY: Oxford University Press.

Zanna, M. P., & Pack, S. J. (1975). On the self-fulfilling nature of apparent sex differences in behaviour. *Journal of experimental social psychology, 11*(6), 583-591.