UNIVERSITÀ COMMERCIALE LUIGI BOCCONI, MILANO
Undergraduate School

# THE IMPACT OF AUDIENCE SIZE ON REPUTATIONAL CONCERNS

Thesis Supervisor: Prof. PIERPAOLO BATTIGALLI

Thesis Author:

SEM MANNA

Student ID 3087964

Submitted to Bocconi University in fulfilment of the requirements for the
Bachelor of Sciences in International Politics and Government
September 2021, Milano

# Acknowldegements

I am most thankful to my Mom and Dad, who put their most valuable resource into the upbringing of me and my siblings: their time. I would be lost without their constant, unconditional love and wisdom. My achievments are theirs as much as they can be mine.

I want to express my gratitude to my supervisor and mentor, Professor Battigalli, for the invaluable support and suggestions, and especially for the words of motivation as well as constructive criticisms dispensed throughout the drafting of this paper. I am also grateful to Professor Stringhi, for the time and kindness devoted to analyzing my doubts and providing crucial reccomendations on how to improve my work. I also owe a special thank you to Emma, for her encouragment and linguistic advices.

Finally, I want to thank Nadir, Yuri and all my friends, for their infinite, contagious fondness which makes me question the cynical foundations of this work.

## Abstract

Why do people care about each other? In an effort to bridge Economics and Evolutionary Psychology, this paper reviews existing explanations of altruism to underline the role played by reputational concerns in supporting prosocial behaviour. The literature on altruism is then applied to strategic interactions to investigate analytically how reputational concerns could be influenced by features of an observing audience. A related experimental design is also proposed to test the theoretical implications of a novel model on size-dependent audience effect. The paper is structured as follows: 1 Introduction and review of the literaturestudies the origins of altruism and image concerns by summarizing the literature in Evolutionary Psychology, and then reviews audience effects in Economics and Psychology to identify a knowledge gap, namely the relationship between reputational concerns and audience size; 2 Modelling audience-dependent reputational concerns elaborates on a model of image concerns by Battigalli and Dufwenberg (2020) to study size-dependent audience effects in a novel Dictator Minigame with external audience; 3 Experimental Design presents an experimental design to test the implications of the model and discusses potential extensions and limitations; finally, 4 Conclusions concludes.

# Contents

# 1 Introduction and review of the literature

## 1.1 Justifications of other-regarding preferences in Economics

Explaining costly prosocial behaviour has been an enduring challenge for economists. Under the abstractions of *Homo Economicus*, altruistic behaviour in rational and self-interested agents must be explained by selfish strategic motivations. To be willing to benefit others at own cost, an individual must believe that the seemingly altruistic act increases her expected utility via future gains in ways which compensate for the immediate loss in material payoff. Hence, among the many formal accounts of seemingly altruistic behaviour, popular solutions to the dilemma have often involved strategic reasoning on reciprocity and gains from cooperation in repeated interactions. Since cooperative equilibria often fall apart when agents cheat, if people believe that past cheaters are more likely to defect again, they will be less likely to be trusted. Then, as forgoing cooperation can be costly, agents should care about their reputations both in the eyes of coplayers and in those of external audiences with whom they may interact in the future.

But analysis of experimental results, most notably by Thaler (1988), showed that individuals are at least partly altruistic even in one-shot, anonymous games with no punishments. Those findings challenged neoclassical economic models while confirming the common-sense belief that humans do indeed care about each other's wellbeing, and that own payoffs are not evaluated in a vacuum, irrespective of those of others.[1]

Other-regarding preferences are now widely employed in economic models and are often included in expected-utility formulations thanks to auxiliary assumptions like "impurely altruistic" motives such as taking pleasure from giving to others (Andreoni (1989)), or social preferences for fair outcomes (Fehr and Schmidt (1999)). The resulting preferences are still selfish in the sense that individuals want to maximise their own utility, but this does not imply that behaviour needs to be selfish. By allowing

[1] The assumption of rationality has also come under scrutiny, but many behaviors can still be accounted for with ameliorations to the rational choice model. See Kahneman (2011) *Thinking, fast and slow* for a collection of biases and heuristics which affect judgment, and Kahneman & Tversky (1979) for a theory of choice under risk which departs from expected utility theory.

the utility of an agent to depend on more than her own material payoff, selfless behaviour can be justified without violating (subjective) rationality and therefore be integrated within adaptations to the neoclassical framework.

Yet, despite accounting for some of the recorded behavioural tendencies, most of these formulations overlook the psychological foundations of human behaviour and consequently overestimate the extent to which we reason strategically, selfishly, and irrespective of what co-players believe or feel. By neglecting human psychology and the evolutionary foundations of the mind, such models may sometime fail to grasp *how* agents are expected to behave in strategic interactions, whilst stopping short of explaining *why* they behave as they do.

The ability of economists to predict human behaviour within the framework utility-maximisation could benefit from a more psychologically accurate approach, while psychology would be enriched by the adoption of formal mathematical models explicating the incentive-driven nature of human behaviour. Psychological Game Theory (PGT hereinafter), which incorporates players' beliefs about actions and beliefs directly into their utility functions, appears to be the most complete effort to bridge the two disciplines in a complete and coherent way suited to model a wide range of human emotions.[2] By allowing to model the dependency of agents' utility on beliefs (or opinions) of others, PGT is a rich and flexible framework to account for reputational concerns and altruism. Finally, appending evolutionary psychology to models in PGT can help develop solid justifications for utility formulations.

## 1.2 The evolutionary foundations of altruism and image concerns

Our understanding of selfless altruism and image concerns, two plausible candidates for explaining most prosocial behaviours, could be improved by detailing the sort of adaptive problems from which they may have originated as evolved solutions.

According to cognitive scientist Steven Pinker, the mind is a system of computational organs forged by natural selection as a design process which develops mental modules with specialised purposes shaped by the sort of adaptive problems faced by our

[2] This subfield of Game Theory was first devised by Geanakoplos et al. in 1989, and later generalised by Battigalli and Dufwenberg in 2009. See also Battigalli and Dufwenberg, *Belief-Dependent Motivations and Psychological Game Theory* (forthcoming).

ancestors (Pinker (1997)).[3] Mental modules work in parallel as parts of an integrated complex to shape thinking, beliefs and desires. Evolutionary psychology and the computational theory of the mind, on which Pinker's 1997 book *How the Mind Works* is based, outline the functions of altruism and reputational concerns by detailing the sort of the adaptive challenges for which they may be evolved solutions.[4]

### 1.2.1 Kin altruism

In modern evolutionary biology, genes – rather than organisms – are considered the replicating unit which compete, evolve, and get selected based on their relative ability to propagate themselves (Dawkins (1976)). In evolutionary terms, the best genes are those giving rise to traits which maximise the number of copies of the genes. This happens when they can pass many copies on to the next generation by increasing the survival or reproductive chances of organisms containing the gene or its copies.

The implications of gene-selection for prosocial behaviour between individuals are major and can explain costly altruism among kin (costly for the individual, but not in terms of the reproductive chances of the gene). Since relatives share part of their DNA sequences, there is a positive probability that they also share copies of a same gene and kin altruism could then be an evolved strategy of genes pursuing the selfish purpose of replicating themselves (Dawkins (1976)). As Pinker explains, when altruism comes at a lower cost than the probability of sharing the gene (or the percentage of DNA shared) times the benefit to the related individual, kin altruism can be an adaptive strategy of naturally selected genes (Pinker (1997), Hamilton (1964a) and (1964b)). Selflessness of this sort can be induced in organisms via valanced emotions which an individual may want to avoid or attain, such as guilt or sympathy.[5] In this view, emotions themselves should be seen as innate, evolved mechanisms to coordinate different and specialised mental modules by defining their respective relative priority to "orchestrate" the working of the mind (Tooby and Cosmides (2005)). Altruistic

[3] It is important to outline, as done by Pinker via a 1992 argument by Donald Symons, that this is does not mean that behaviours are themselves directly adaptive or prescribed by evolution.
[4] The idea of applying evolutionary principles to the mind is not new. Charles Darwin anticipated it in *On the Origin of Species* (1859): "Psychology will be based on a new foundation, that of the necessary acquirement of each mental power and capacity by gradation". William James also mentioned "evolutionary psychology" explicitly in *The Principles of Psychology* (1890).
[5] See also James (1884) for a discussion on what emotions are and Gazzaniga et al. (2014), Ch. 10, for different perspectives and the cognitive neuroscience of emotions.

emotions should then be felt stronger, for instance, the better the condition of the altruist, the worst those of the recipient, the smaller the cost of the act, and the stronger the blood ties between the two.

Consequently, kin altruism has been exploited as a potential justification for altruistic behaviour among unrelated individuals. If prehistorical humans lived in small kinship-based groups of close relatives, selfish genes may have induced some pure and indiscriminate altruistic behaviour. In today's vast societies this would then translate into the maladaptation of extended altruism towards unrelated individuals in one-shot encounters.

Although kin altruism is validated by psychologists, the theory on its extended consequences is more debated. According to moral psychologist Jonathan Haidt (2012) it should be completely ruled out, as it is falsified by anthropological evidence, notably from Hill et al. (2011). According to their work, 90 percent of group mates in societies of hunter-gatherers were unrelated individuals or distant kin, and consequently the sort of indiscriminate altruism posited could not have evolved. Kin altruism is therefore an unlikely candidate for explaining extended prosocial behaviour.

### 1.2.2 Reciprocal altruism

Once kin altruism is ruled out, most of the remaining evolutionary explanations of altruism rely on reciprocity and the gains from cooperation that it can produce for the individual.

Repeated cooperative interactions, by producing multiple streams of large payoffs, can produce considerable benefits for the agents (and genes) that manage to establish them, making them profitable regardless of considerations of kin. Delton et al. (2011) believe that repeated gains from cooperation are so relatively large when compared with the costs of a single-shot exploitation, that human altruism may be justified as an evolved mechanism to take risks aimed at establishing dyadic cooperation under uncertainty. According to the authors, even if losses from cheating can be expected, altruism evolved because the gains from the repeated, reciprocal partnerships that it could lead to are large enough to compensate for some risk of exploitation.

Yet some free riders – who exploit altruistic agents and never reciprocate – may soon overcome all the altruists in the population. According to biologist Robert Trivers' theory of reciprocal altruism (1971), for costly altruism to survive within a population free riders must be identified and then excluded or punished. In Trivers' model, reciprocation can motivate altruism on the basis that favours will be returned in the future but, in absence of enforcement mechanisms, cheaters will be better off than altruists, and free riders will eventually dominate the population. For altruism to survive then, cheater-detection mechanisms must evolve to spot, remember, and punish free riders (or reciprocate favours only with other altruists). In turn, cheaters evolve to disguise non-reciprocation or minimise contributions, and spotting systems respond by becoming more and more complex in a "cognitive arms race" between the two (Pinker (1997)). In humans, this gives rise to complex systems of strong moralistic emotions which are hard to fake, making promises and costly threats ex-ante credible and thus supporting cooperation by means of reciprocal altruism.

As underlined by Pinker, reputations become extremely valuable in a system of reciprocal altruism and attempts to gain trustworthiness and identify cheaters drive the insatiable human thirst for gossip. He also points out how evidence from evolutionary psychologist Leda Cosmides (1985) found cheater-detection mechanisms in humans. In her experiments, individuals were more capable of solving (identical) logical problems when they were framed as exchanges of benefits where cheaters had to be found (rather than as plain logical questions). The identification and communication of cheaters (or gossip) may therefore be the backbone of human reciprocal altruism. This could partly explain why people are obsessed with what others think of them, especially friends and relatives with whom they interact constantly. Audience concerns are a direct extension of this principle.

Identifying cheaters supports reciprocal altruism but may not always be sufficient to produce cooperation. Reciprocating individuals should also be able to identify each other, and altruism can serve as a signalling strategy to communicate one's suitability as a cooperation partner. This is what *indirect reciprocity theory* maintains (Alexander (1987)), and the reputational benefits may extend beyond recipients as information spreads through gossip (Buss (1998)). According to the theory, indirect reciprocity

could also justify pure altruism towards strangers in the absence of direct observers. This holds true if altruists can benefit from cooperative interactions – and hence improve their survival and reproductive chances – by means of reputations gained by word of mouth, regardless of who initially benefited from the altruistic gesture (Buss (1998)).But not all reputational partners or audiences should be valued the same. Due to their differing ability to mobilise resources and partners, influence reputations within the social network, and more generally to undertake actions that can be expected to impact others, different weights should be placed on different observers. Moreover, different onlookers may hold different preferences, beliefs, and expectations and thus evaluate actions differently, with considerable consequences on reputations. The nature of their interaction with different groups may also vary, and agents may wish to attain different reputations with different audiences. Therefore, actions may interact with the audience's beliefs and preferences in order to produce valued reputations.

But regardless of how they may be formed, inferred, or evaluated, reputations are recognised as central in all theories of reciprocal altruism. Yet, most do not specify how individuals infer their social image. The *Sociometer theory* (Leary and Baumeister (2000)) could close this gap. According to the authors, self-esteem is evolutionarily developed as an internal mechanism to gauge one's own value as a relationship partner. Selfless behaviour towards strangers in absence of observers could then be vindicated as a by-product of said internal indicator, meant to keep us informed of our likability and eager to increase it (via means of unpleasant warnings or amiable validations).

According to Steven Pinker, theories of reciprocal altruism are backed by anthropological evidence, especially from the Aché indigenous people in Paraguay and the !Kung people from the Kalahari Desert. In both groups, where people seem incredibly selflessly altruistic towards one another, the goods shared with others are in fact only those are attainable with high variability and can hence produce the largest gains from favour-exchanging (Pinker (1997)). Favours can then be seen as a mean to "store" surpluses of goods available with high variance in the form of a special credit of favours to be reciprocated in times of need. The human ability to develop mental ledgers to record and remember favours exchanged seems coherent with this view.

Accordingly, reciprocal altruism should also be regarded as a form of risk-pooling among individuals designed by natural selection. Petersen et al. (2012) further found that the same holds for goods that are obtained by luck as compared to those acquired by effort, which are shared less to avoid free riding. Probabilistic events producing large payoffs – such as a successful stag hunt – generate larger benefits from reciprocal sharing at lower risk of free riding with respect to effort-dependent activities such as gathering food.

All the aforementioned explanations of altruism, grounded on reciprocity, highlight the importance of reputation and the related audience concerns, but fall short of explaining all altruistic behaviour in anonymous interactions with strangers. Pure altruism may have likewise evolved thanks to selection pressures between groups.

### 1.2.3 Group altruism

In *The Righteous Mind* (2012), Jonathan Haidt considers evidence for the emergence of moralistic feelings and reasoning which completes Pinker's analysis of individual- or gene-level selection during the Pleistocene with multilevel selection and more recent evolution. Haidt does so by accounting for possible evolved traits from selection pressures at the level of groups, including those that according to modern genetic evidence may have occurred in the last few millennia.[6]

Group selection was already theorised by Charles Darwin in 1871, and multilevel selection analysis allows to quantify evolutionary pressures at multiple levels to justify the emergence and survival of genes adapted in response to them (Haidt (2012)). On group selection, Darwin argued that tribes of sympathetic individuals could arrange organised violence and superior warfare to dominate others (Darwin (1871)). Self-sacrifice and altruism can support cooperation and hierarchical relationships within the group consequently allowing for improved resource production and coordination of war-related efforts. More organised and *groupish* tribes, by means of faster growth, conquest, and annexation would likely dominate less cohesive tribes and shape selection pressures at the level of groups. But for cohesive groups to emerge,

[6] Theories on group selection and adaptive pressures in the last few millennia do not run counter to what argued by Pinker. In fact, they could be included in his analysis to enrich it, as stated by Pinker in the 2009 foreword of *How the Mind Works*.

individuals within them must seemingly defy individual-level selection to cooperate at potentially massive individual risks (such as injury or death from combat). To avoid allowing free riders to jeopardise cooperation, tribes must develop enforcement mechanisms to punish and quell selfishness, and their ability to do so determines their evolutionary success over other groups. According to Haidt's summary of Darwin, this could have contributed to the emergence and strengthening of emotions associated with image concerns, norms-violation and human ultra-sociality (Haidt (2012)).

Human history seems to be coherent with this narrative. Jared Diamond (1997) argued that kinship societies in the Pleistocene were already violent and belligerent before the invention of agriculture in the Neolithic.[7] Warfare subsequently surged because of agricultural surplus, demographic growth, and the emergence hierarchical societies able to conquer and assimilate others; but selection pressures among groups may have been present before, for periods of time more significant for evolutionary processes. If competitive adaptive pressures at the level of groups contributed to shape the human mind, people would not only be selfish (and clannish), but also *groupish* (Haidt (2012)).[8]

Drawing from multilevel selection, the relative strength of evolutionary pressures at the level of groups should be a function of the amount of warfare and competition between them.[9] When warfare increases, perhaps due to demographic tensions and environmental or social circumscription, victorious primordial polities dominate or assimilate the others (Carneiro (1970)) and group selection should become more prominent. This has been the case starting from when the Neolithic revolution, 12,000 years ago, allowed for the accumulation of agricultural surplus and the emergence of stratified political entities. If larger and hierarchical social structures produced ubiquitous warfare and increased group-level selection pressures, whether to accept

[7] See Ember et al. (1997) for a cross-cultural analysis of past violence, war and aggression; Keeley, L. H. (1996) for a dismissal of the *bon sauvage* and peace among kinship societies; and Glowacki et al. (2017) for a study on warfare from the perspective of evolutionary anthropology.

[8] For an explanation of groupish behaviour without relying on group selection, see Tooby and Cosmides (2010).

[9] More precisely, group-level competition is more than just warfare. Ultimately, in evolutionary terms groups are competing in their ability to use resources to produce surviving offspring (see Lesley Newson in Haidt (2012)).

group-selection or not rests only on the length of time required, in evolutionary terms, to produce significant adaptations.

Haidt believes that a few millennia could be sufficient. By reviewing experimental evidence on animals, he shows that significant evolutionary adaptations can happen in few generations under extreme selective pressures (forced by the experimenter). More notably, Haidt argues that results from DNA analysis carried out by the Human Genome Project show an acceleration of evolutionary processes during the last 50,000 years and especially in the last 20,000 years (Haidt (2012)). In Tibet, for instance, people have developed genes that increase their resistance to high altitudes, and since the domestication of mammals, humans have developed lactose tolerance (Haidt (2012)).[10] If the mind evolved like other biological structures, ultra-sociality and groupishness could be a recent adaptation of existing mental modules initially developed to manage kin and reciprocal altruism to pressures from group-selection.

Along with altruism, cooperative behaviours beneficial to the group could have emerged and been enforced through social norms, thanks to mental programs allowing individuals to identify and comply with them. To cooperate, two or more minds must be equipped with the ability to share the *intentionality* and the understanding of problems that could be solved by cooperation (Tomasello et al. (2005)). Primordial mental modules which allowed humans to understand their ability to solve survival problems through dyadic cooperation could have then adapted or found new purposes in larger social environments. Here, they survived and evolved more complex cheater-detection and enforcement mechanisms, limiting free-riding to produce larger payoffs at limited costs or risks. Subsequently, with the domestication of plants and animals and the emergence of more stratified and belligerent societies, group-level pressures probably gave the upper hand to groups which developed more complex forms of large scale, extended cooperation and *groupishness*.

This process results in behavioural expectations grounded on prescriptive rules – or social norms – among group members which individuals feel (also emotionally) bound to. Once established, social norms further reinforce cooperation via selection pressures

[10] See Pickrell et al., 2009 for more evidence on recent adaptations.

within groups which favour more collaborative members over selfish ones – who may be punished or excluded from cooperative ventures (Haidt (2012)). Haidt further argues that social norms (and culture in general) could have coevolved with mental and biological traits (shaped by genes) to allow humans to feel moralistic emotions.[11] At the group level, natural selection favoured internally peaceful, cooperative, and stratified groups over disorderly tribes of selfish individuals. In turn, cooperative groups developed more complex societies and institutions (including religion) which strengthened the selective pressure of their members in ways which favoured norms-compliance and strengthened their culture. The resulting selective cycle favours somewhat cooperative group members which comply to social norms on one side, and stratified cultures which can organise food production and warfare more efficiently on the other.[12] Moralistic emotions and *groupishness* could therefore be the result of selective pressures from groups over individuals, emerging as a consequence of competitive pressures between groups.

In that case, humans should have evolved mental devices making prosociality and norm abidance rewarding for the individual, regardless of considerations of kinship and reciprocity. The application of neuroimaging devices to economic experiments seems coherent with this view. A review of the scientific literature by Fehr and Rockenbach (2004) shows that cooperation and punishment of defectors produce neural responses associated with reward circuits, thus making altruism psychologically pleasant for humans.

But for selfless altruism and a *groupish* mentality which supports cooperation to be stable in evolutionary terms, prosociality should evolve alongside traits allowing humans to avoid positive spillovers to other groups. Experimental evidence from Bernhard et al. (2006) shows exactly this: human cooperation is sustained by social norms and "parochial altruism", which selectively favour members of one's own community or ethnic group over outsiders.

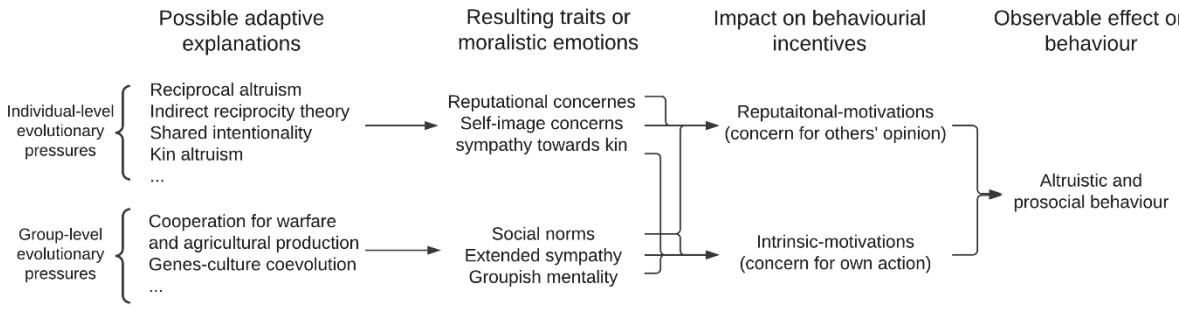
[11] Fehr & Fischbacher (2003) argue that some relevant altruistic behaviours in humans cannot be accounted for by gene-based evolutionary theories alone and illustrate the importance of theories based on cultural evolution and gene-culture coevolution.

[12] The selection of cooperative (or compliant) individuals, by groups and through reciprocity and norms enforcement mechanisms, is called "self-domestication". More on this in Hare et al. (2012).

### 1.2.4 Recap: The evolutionary foundations of altruism and image concerns

As argued through the literature summarised, human altruism is probably the consequence of an evolutionary process shaped by selective pressures at the level of individuals and – to a smaller extent – groups arising respectively from reciprocal cooperation and conflict. Those evolutionary adaptations include forging and shaping mental modules to motivate altruistic behaviour via intrinsic motivations and reputational concerns about the opinions of others. The recapitulatory diagram in Figure 1 summarises these claims. To recognise this argument as valid is to say that human behaviour is influenced by incentives coming from valanced emotions shaped by evolutionary adaptations.

*Figure 1 – recapitulatory diagram: evolutionary foundations of altruism and image concerns*




## 1.3 Audience effects and observability in Psychology and Economics

### 1.3.1 Social facilitation and audience effects

Observability is key for reputational concerns to influence behaviour, and the effects of being watched are among the first topics studied by psychologists. In one of the very first laboratory experiments in Social Psychology, Triplett (1898) studied how children's performance in physical efforts improved in the presence of a counterpart. The effect of observers on individual performance is now well known in Sport Psychology under the name of *social facilitation*, and its implications extend beyond performance and sports, permeating all fields of Psychology.

In 1965, Robert Zajonc distinguished social facilitation between *audience effects*, coming from the mere presence of passive spectators, and *co-action effects*, arising from the presence of other individuals engaging in the same activity. Subsequently, experimental evidence on social facilitation by Cottrell et al. (1968) showed that active spectators influenced students' performance in verbal tasks to a larger extent than

passive ones. Their work pointed to the effects induced by the awareness of observers as active, watching agents, as opposed to the effects induced by the mere presence of non-watching individuals. Considering this, psychologists now define audience effects as behavioural changes "caused by being observed by another person, or the belief that one is being observed [...] This requires [...] some level of awareness that the other is watching, that is, awareness of the perceptual state of the other" (Hamilton and Lind (2016)).

Audience effects can include a range of behavioural responses, including pro-sociality, through the impact of observers on reputational concerns. This has been documented extensively across a wide range of disciplines. Most relevantly, a meta-analysis of 90 experimental studies from Abeler et al. (2019) suggests that, along with intrinsic motivations for being honest, people also care about being perceived as such, and inducing observability results in higher frequency of truth-telling (similar results in Fischbacher and Föllmi-Heusi (2013)). Reputational concerns also motivate people to engage in prosocial behaviours (Bénabou and Tirole (2006); Ariely et al. (2009); Cañigueral and Hamilton (2019)) such as contributing to public goods (Ozbay and Ozbay (2013)), volunteering (Linardi and McConnell (2008)), donating to charities (DellaVigna et al. (2012); Soetevent (2011)) and can shape firm behaviour (Sebald and Vikander (2018)) or sport performances (Böheim et al. (2019)).

### 1.3.2 Observability in Economics

The literature in Economics further shows how strongly reputations can influence behaviour to produce conformity and induce prosociality when individuals are being observed. A theoretical model by Bernheim (1994) predicts that individuals with private and unobservable traits, who care about the public perception of such traits, may repress their individuality to display conformity to a social norm. Subsequently, by drawing on puzzling empirical evidence, Andreoni and Bernheim (2009) argue that recorded altruism and preferences for equal splits alone cannot explain key experimental findings. Thus, they propose and test a theoretical framework for studying audience effects grounded on the assumption that people also wish to be perceived as fair. In their model, fairness is displayed by complying to a social norm – that of equal splits. Consequently, individuals may be willing to give up some

material payoff up to the point where payoffs are shared equally to signal their own fairness and achieve a desired social image. When agents believe that a social norm is in place, signalling equilibria may emerge and deviations from the norm could become informative signals of private traits such as (lack of) fairness. If some altruism is expected, seeming altruism may then be, to some extent and for some actors, a signalling strategy of agents with reputational concerns to display private traits such as fair-mindedness (or fitness for reciprocity and cooperation in evolutionary terms) via norm-compliance. Andreoni and Bernheim (2009) tested their theory in a laboratory experiment involving a Dictator Game and found support for the claim that people care about being perceived as fair. Their findings are also consistent with empirical evidence on how monetary incentives can fail to promote prosocial behaviour. Analysed in their framework, payoff compensations for prosocial behaviours could crowd-out reputational motivations by diminishing the signalling power of altruistic actions, thus potentially decreasing the frequency at which they are carried out. One research on monetary incentives and observability, carried out by Ariely et al. (2009) and based on novel laboratory and field evidence, shows that introducing material incentives is indeed ineffective at motivating prosocial behaviour when actions are publicly visible.

Attanasi et al. (2019) also studied the presence of image concerns experimentally. Their study shows that graduate students playing a public good game behave differently when their actions are observed by future cohorts of students (without payoff linkages between the two). Surprisingly, the students did not seem to wish to signal their fairness, and when their actions and photos were transmitted to future participants, they contributed significantly less. The authors still attribute this behaviour to image concerns, supposing that students wish to not be perceived as "*suckers*". In this case, audience characteristics – such as age or the social environment – may have had a considerable impact on the outcome in ways that defied common expectations. An interpretation of this could be that there is more to reputations than just altruism or fairness, and that individuals wish to manage their reputation from different audiences strategically. In a competitive and socially exposed environment such as universities, students may wish to signal toughness over

altruism. This could in turn be grounded on sexual selection, as potential mating partners who find this behaviour attractive could populate the audience for experiments carried out among university students.

Ozbay & Ozbay (2013) also studied the effect of third-party audiences on efforts to produce public goods. In their experiment, once external spectators observing efforts to produce public goods are introduced, free riding falls significantly and people contribute more. The same effect is absent when contributions to the public good are not linked to efforts, or when efforts produce private returns, meaning that there is no payoff or strategic linkage between players. As expected, the experimental results show that observability is not relevant when players' actions do not affect the payoffs of others or signal private traits (Ozbay & Ozbay (2013)). These results further show that agents care about their reputation as hard workers for their group. Evolutionary explanations grounded on reciprocity, reputations and social norms, and group selection may explain the observed behaviour.

With few exceptions, such as in Andreoni et al. (2019) and Ozbay and Ozbay (2013), most experiments on observability in economics study audiences comprised of coplayers. That is, they focus on identifiability from co-players who are involved in the game and may thus be directly relevant for strategic or psychological reasons that go beyond reputational preferences of external agents unrelated to the ongoing strategic interaction. For instance, in public good games with observability, larger contributions could be motivated by reciprocity or the desire to establish a norm for cooperation, rather than by image concerns. This is supported by experimental evidence from De Cremer and Bakker (2003). Their work shows that non-anonymous agents playing social dilemmas contribute more when they believe that coplayers have image concerns that incentivise cooperation on their side. This has to do with the idea that reputational concerns of others within the group display their awareness of a social norm, and non-cooperation will therefore be judged negatively. Contributions could then also be interpreted as signalling strategies to display own beliefs over the existence of a norm for cooperation, hence incentivising cooperative behaviour via image concerns.

Possibly, the role of external audiences has been partly overlooked because, under the prescriptive neoclassical framework, they should be irrelevant for rational and self-interested agents. On the contrary, field evidence from a range of disciplines shows that audiences do influence decision-making. Yet, to properly understand audience effects, they should be studied empirically and experimentally with audiences comprised of external, inactive observers with fixed payoffs and no stake in the outcome of the game. Moreover, the strength of audience effects may vary with its size and other characteristics which should be accounted for. To the best of my knowledge, no economic model accounting directly for size-dependent audience effects has ever tested in an experiment involving external spectators.

### 1.3.3 Knowledge gap

Despite a growing corpus of articles in economic literature to account for reputational concerns, most theories do not explicit to the reader the psychological or evolutionary motivations underpinning them. Moreover, external audiences have been often overlooked, along with size-dependent reputational concerns.

This work contributes to the literature by studying how the size of an external, inactive audience without any payoff linkage with players could influence behaviour in strategic interactions via reputational preferences. Possible evolutionary interpretations have been introduced and discussed lengthily, and an experimental design and the related game-specific utility functions are proposed to test the implications of the theory.

### 1.3.4 Recap: Audience effects and observability in Psychology and Economics

Evidence from Economics and Psychology seems to support findings and assumptions of evolutionary psychologists: altruism emerges from a mix of intrinsic and extrinsic motives driven by reputational concerns and grounded on the observability of actions. Yet, most work in Economics focuses primarily on concerns over the opinion of strategically relevant coplayers.

This work, by developing and studying a size-dependent model of audience effects mediated by reputational concerns over the opinion of external observers, aims at complementing the existing literature while providing well-founded interpretations for audience effects grounded in evolutionary psychology.

# 2 Modelling audience-dependent reputational concerns

## 2.1 Theoretical framework and formalism

To account for audiences' size and features, I model image concerns within the mathematical framework of PGT put forward by Geanakoplos et al. (1989) and later advanced by Battigalli & Dufwenberg (2009). Belief-dependency, and hence PGT, is appropriate to properly study audience effects, for they impact behaviour through reputational concerns, modelled as the dependence of psychological utility on the beliefs of others over own private traits. The formal methodological framework follows the forthcoming survey by Battigalli & Dufwenberg (B&D henceforth) and the 2019 article by Battigalli, Corrao, and Dufwenberg.

Before moving to the models, I briefly introduce in Table 1 a few concepts and notations coming from B&D which will be used throughout the rest of the paper:

*Table 1*

| | |
|---|---|
| $j \in I$<br>$i \in I$ | players, including nonactive ones, are denoted by $\boldsymbol{j}$. $\boldsymbol{I}$ is the finite set containing all players. When referring to players' own beliefs, traits, actions, utility, and payoffs, the letter $\boldsymbol{i}$ is used |
| $h \in \bar{H}$ | a history $h$ describes a unique sequence of action profiles in the set $\bar{H}$ of possible histories |
| $z \in Z$ | $z$ is a terminal history in the set of possible terminal histories $Z \in \bar{H}$. The game ends here as payoffs and terminal beliefs are realized |
| $\mathrm{a}_i \in A_i(h)$ | $\mathrm{a}_i$ is an action of player $i$ which belongs to the set of her feasible actions $A_i$ at history $h$ |
| $\alpha_i \in \Delta_i^1$ | a first-order belief $\alpha_i$ is a (subjective) probability measure on how the game will be played. The space of 1st-order belief systems of player $i$ includes all the first-order beliefs $a_i(\cdot \mid h) \in \Delta(Z)$ which $i$ deems possible conditional on observing history $h$ |
| $\beta_i \in \Delta_i^2$ | a second-order belief $\beta_i$ is a joint belief over paths of play (actions) and first-order beliefs of other players. Second-order beliefs are relevant if players' utilities depend on the first-order beliefs of others. A first-order belief system can be derived by marginalization from a second-order one since it is included in it |
| $CB(E)$ | common belief of $E$. It means that everyone is certain of $E$ and of the fact that everyone else |

| | |
|---|---|
| | is too. Everyone also believes that that everyone else thinks that all others are certain about $E$ too. Everyone is also certain that everyone believes this too, and so on |
| $\pi_i: Z \to \mathbb{R}$ | $\pi_i$ is the material payoff function of player $i$. It maps terminal histories to real numbers which are material (monetary) payoffs |
| $\boldsymbol{\theta}_i \in \Theta$ | $\boldsymbol{\theta}_i$ is a vector of private traits of $i$ which affects her utility. Players are heterogeneous over their traits |
| $u_i: \Theta_i \times Z \times \Delta^1_{-i} \to \mathbb{R}$ | $u_i$ is the psychological utility of player $i$. In this class of psychological games, it depends on her private traits, the terminal history reached, and the set of conditional first-order belief held by all other players over actions and unobservable traits of $i$ |

Finally, in the model players are assumed to plan rationally. Following B&D, rational planning is satisfied for a belief system $(\alpha_i, \beta_i)$ when incentive-compatibility holds, meaning that a player $i$ assigns positive probability only to own actions that maximize her expected utility $\bar{u}$ at time of the choice, and are therefore local best replies:

$$\alpha_{i,i}(\mathrm{a}_i|h) > 0 \Rightarrow \mathrm{a}_i \in \arg \max_{\mathrm{a}'_i \in A(h)} \bar{u}_{i,h}(\mathrm{a}'_i; \beta_i) \qquad 1$$

Moreover, the belief systems of cognitively rational players should be updated according to the rules of conditional probability whenever possible.

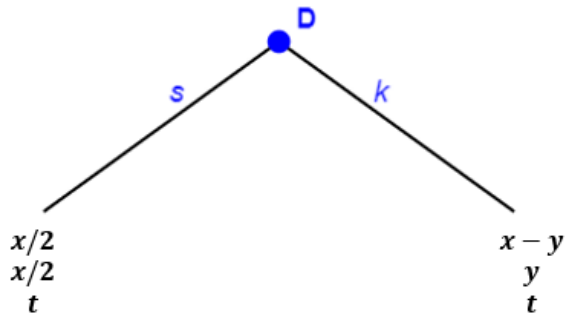
## 2.2 The Dictator Minigame with external audience

To study the interaction between reputation-building and external observers, I propose an experimental design which involves a simple variation of the Dictator Minigame to introduce an external audience of variable size (external meaning in addition to the internal audience composed by the co-player). The game form involves an active player – the Dictator (D henceforth) – and a passive one – the Recipient (R).[13] At the root of the game (the empty history $h = \emptyset$) D is endowed with an amount $x$, and can make a dichotomous choice between *keep* $(k)$, which yields material payoff $\pi_D(k) = x - y$, with $y$ smaller than $\frac{x}{2}$, leaving $\pi_R(k) = y$ to R; and *split* $(s)$ which

[13] Neutral wording will be used in the actual experiment to avoid confounders.

leaves both players with an equal payoff of $\pi_D(s) = \pi_R(s) = \frac{x}{2}$. Since $y$ is smaller than $\frac{x}{2}$, the material payoff of Receivers is larger following action *split*, while the opposite holds for Dictators. If present, members of the external audience are passive (pseudo) players whose payoff $\pi_A = t$ is unaffected by the terminal history ($z$) realized. All players perfectly observe $z$, and can therefore infer the unique action ($a_D \in \{k, s\}$) chosen by D. The game form is summarized through the game tree in Figure 2.

*Figure 2 – The Dictator Minigame with external audience*




The aim is to investigate whether observation from inactive players, without payoff linkages nor future interactions, can promote prosocial behaviour via psychological incentives for being perceived as having high intrinsic motivations for performing good actions. An additional question is how, if present, such effects vary in the size of the audience.

In the following sections, utility formulations to account for size-dependent image concerns are proposed and then applied to the Dictator Minigame to study the behavioural implications of changes in the presence and number of spectators. The utility of players in the role of the Dictator is assumed to depend, other than on material payoffs, on intrinsic motivations for undertaking good actions and on their reputation, which depends on the material payoff left to R following their action and on the size of the observing audience. Other-regarding emotions which could influence D's behaviour, such as guilt, are ignored as they should be unchanged in the presence and size of the audience.

## 2.3 Reputational concerns about intrinsic motivations

Image concerns are modelled after B&D. In their formulation, agents are moved by reputational concerns as well as intrinsic motivations to undertake "good" actions.

According to B&D, the strength of selfless motives depends on a private trait which is heterogeneous in the population and whose perception in the eyes of others gives rise to reputational concerns. Consequently, even selfish agents with strong reputational concerns may have incentives to behave in ways which influence other's *ex-post* perception of their private trait, and this type of psychological utility induces signalling. This theorization is consistent with the sort of intrinsic and extrinsic (reputational) behavioural incentives found by evolutionary psychologists and discussed in *1.2 The evolutionary foundations of altruism and image concerns*. I use their utility formulation to discuss and study experimentally how the number of spectators could impact on reputational concerns. Image concerns about intrinsic traits are introduced via a generic utility function and later applied to the Dictator Minigame to capture size-dependent audience effects in games:

$$u_i(z, \alpha_{-i}, \theta_i, n) = \pi_i(z) + \theta_i^I\left[I_i^G(z) - I_i^B(z)\right] + \sum\nolimits_{j \neq i} \theta_{ij}^R \mathbb{E}_{\alpha_j}[\tilde{\theta}_i^I | z] \qquad 2$$

As usual, players are motivated by material payoffs, $\pi_i(z)$, while the second term in (2) accounts for intrinsic motivations to do good deeds, parametrized by a private trait $\theta_i^I$. The indicator function $I_i^G(z)$ takes value 1 if $i$ carried out subjectively "good" actions in path $z$, and 0 otherwise. The opposite holds for $I_i^B(z)$ and their difference measures the net intensity of the "goodness" of $i$'s actions in path $z$.[14] As accounted for in the third term, an agent also cares about her reputation, which here depends on what others expect her intrinsic-motivation trait to be at the end of the game. Since players cannot observe the beliefs of others, this information is private and probabilistic inferences over others' opinions are formed by consulting one's own (conditional) second-order beliefs.

Reputations are then the expected value, based on the system of second order beliefs of $i$, of the expected value of her intrinsic trait in the eyes of the observers, based on their first-order beliefs and conditioned on the end node reached. Therefore, when

[14] The "goodness" of an action profile may be subjective, but I think it's safe to assume that in a constant-sum game (in terms of material payoffs) like the Dictator Minigame and with roles assigned randomly, both players consider each other *ex-ante* equally worthy of receiving money. Therefore, $s$ is always and unambiguously regarded by everyone as a "good action", while $k$ is regarded as a "bad" one. I took inspiration from B&D and Andreoni and Bernheim (2009).

writing the utility of actions of D in the Dictator Minigame, $j$'s expectation of the intrinsic trait of D ($\mathbb{E}_{\alpha_j}[\tilde{\theta}_D^I|z]$) will be expressed as $\mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_j}[\tilde{\theta}_D^I|z]\right]$. The expectation of D is based on her initial SOB, and therefore not conditioned on any history, because she is the only active player and her evaluation of her psychological utility at the root of the game leads to an action mapping directly to outcomes.

More notably for studying size-dependent audience effects, in B&D's formulation the overall magnitude of a player's reputational concern depends on the linear combination of all the dyads of image concerns between the player and any possible coplayer. In other words, $i$'s overall concern for the opinion of others depends on the sum of their individual opinions weighted by how much she cares about the opinion of each. The trait vector of player $i$, $\boldsymbol{\theta_i} = (\theta_i^I, \boldsymbol{\theta}_i^R)$ contains her intrinsic-motivation trait $\theta_i^I$ and the vector $\boldsymbol{\theta}_i^R$ of all the reputational-motivation traits associated with every coplayer $j$ of $i$.

Assigning different decisional weights to the opinions held by different players may have both strategic and evolutionary justifications. On the strategic side, for instance, some players may have a higher status, and hence be more capable of mobilizing resources, knowledge, or other people to attain some goal. Their opinion could then be regarded as more relevant, as they can be potentially more rewarding cooperation mates. Others' past behaviour in reciprocal cooperative relationships should also be a factor, as it is a meter for the probability of establishing successful collaboration. Moreover, as players communicate and information spreads through gossip beyond the initial audience, some players may be more capable of influencing the opinions of others by virtue of their individual ability, position in the social network, the strength of their ties with others, and more. This, in turn, influences one's ability to establish cooperative relationships. All these factors, and many more which have not been listed here, should influence the weight assigned to the opinion of a spectator for strategic reasons. On the evolutionary side, relevant individual characteristics of spectators could include kin, clan, religious beliefs, race, age, biological sex, past cooperation history, reputation in the eyes of others, frequency of involvement in gossip, or norms-

abidance.[15] Not surprisingly, some evolutionary justifications overlap with the strategic ones, as should be expected if strategic gains translate into selective advantages.

### 2.3.1 Simplifying assumptions

For the purpose of this work, $\theta_i^I$ and all elements of $\boldsymbol{\theta}_i^R$ are assumed to be nonnegative, meaning that people do not enjoy performing "bad deeds" or having a "bad" reputation. Moreover, this is assumed to a be commonly held belief among players.

$$\theta_i^I \geq 0 \;\&\; \theta_{ij}^R \in \boldsymbol{\theta}_i^R \geq 0;\; CB\left(\theta_i^I \geq 0 \;\&\; \theta_{ij}^R \in \boldsymbol{\theta}_i^R \geq 0\right) \qquad 3$$

Under those assumptions it is easy to derive from (2) that any additional audience member adds a nonnegative term to $i$'s reputational concerns, and audience concerns are therefore monotonically increasing in the size of the audience. Consequently, reputations can only induce an increase in utility from having a "good" image. If reputations cannot be negative, increasing the audience size can only increase one's utility, or leave it unchanged at worst.[16] While being normatively questionable, this should not change how I expect audiences to influence behaviour in the Dictator Minigame: rather than having psychological costs for bad reputations and benefits for good ones, all effects of image concerns are captured via the increase in utility from a good reputation.

To further simplify the analysis and derive the impact of the number of observers through some intuitive comparative statics, I make the additional assumption that all other players are identical from the standpoint of $i$. This simplification should not compromise the analysis and is justified by the anonymous setting of the experimental design where reputational concerns will be tested. The assumption implies that, from

[15] A disclaimer may be necessary here. To say that the mind may have evolved to prefer, for instance, sexual partners or more (seemingly) genetically similar individuals over others does not mean to that sexism, nepotism, or racism should be condoned. One should always be wary of the "naturalistic fallacy" and, as remarked by Pinker (1997), we should separate "good biology" and "good ethics", with each being autonomous from the other.

[16] An example applied to the Dictator Minigame may help clarify this logic. If an increase in $n$ induces a D-type who would have otherwise played $k$ to play $s$ instead, it must that the utility from a "good" reputation plus – unchanged – intrinsic motivations now outweigh the – unchanged – material incentives to play $k$ and give up the psychological benefits of a good image. D's utility has then increased! If instead D still prefers to *keep*, her utility has remained unchanged as, since only "good" reputations matter, an increase in $n$ does not pose any psychological cost for playing $k$ and material incentives still dominate the image benefits of playing $s$.

the standpoint of a player $i$, $\mathbb{E}_{\beta_i}\left[\mathbb{E}_{\alpha_j}[\tilde{\theta}_i^I|z]\right]$ and $\theta_{ij}^R$ are both equal across all and every observer $j$ in $I$. This means that, while $i$ may not have a deterministic belief over what others think her intrinsic-motivation trait is conditional on the terminal history reached, her expectation of their guess is identical for each observer. Moreover, for $i$ and taken individually, the opinions of all other players are equally important. Hence, $\theta_{ij}^R$ and $\mathbb{E}_{\beta_i}\left[\mathbb{E}_{\alpha_j}[\tilde{\theta}_i^I|z]\right]$ can then be replaced, respectively, by a parameter $\theta_i^R$ and a conditional expected value of the intrinsic trait, $\mathbb{E}_{\beta_i}\left[\mathbb{E}_{\alpha_{-i}}[\tilde{\theta}_i^I|z]\right]$, equal for all coplayers of $i$.

$$\forall j \neq i \in I, \theta_{ij}^R \equiv \theta_i^R \;\&\; \mathbb{E}_{\beta_i}\left[\mathbb{E}_{\alpha_j}[\tilde{\theta}_i^I|z]\right] \equiv \mathbb{E}_{\beta_i}\left[\mathbb{E}_{\alpha_{-i}}[\tilde{\theta}_i^I|z]\right] \quad 4$$

As shown in the following sections, this assumption allows me to use a single set of possible inferences over $i$'s trait and a single reputational parameter for all audience members both in the case of linear (5) and radical (8) dependency of reputational concerns on the audience size.

### 2.3.2 Linear dependency: application to the Dictator Mini-Game

The simplifying assumption in (4) reduces the overall magnitude of reputational concerns of a player $i$ in B&D's formulation (2) to the simple product of the number of observers times the reputational weight of any coplayer and whatever their guess over $i$'s intrinsic trait is (which, being their private information, $i$ infers through her second-order beliefs):

$$\sum_{j \neq i} \theta_{ij}^R \mathbb{E}_{\alpha_j}[\tilde{\theta}_i^I|z] = n \cdot \theta_i^R \mathbb{E}_{\alpha_{-i}}[\tilde{\theta}_i^I|z] \quad 5$$

Applied to the Dictator Minigame, $-i$ is the finite set of spectators observing the game and it comprises all active and inactive players, including the external audience, omitting $i$. By assumption (4), for D all audience members are identical and each action $\mathrm{a}_D$ leading, in the Dictator minigame, to terminal history $\mathrm{a}_D$ is expected to produce a reputation of $\mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}[\tilde{\theta}_D^I|\mathrm{a_D}]\right]$, equal for all audience members. As stated, reputational concerns will be expressed through the expectation, based on the SOB of D, of the audience's guess, every time the psychological utility of actions are being calculated or compared.

Moreover, the resulting game (obtained when preferences appended to the game form) induces signalling: in the Dictator Minigame, action $k$ signals a low intrinsic motivation trait $\theta_D^I$ and therefore, for reputational trait $\theta_D^R$ high enough, even selfish Ds may decide to play $s$ to induce the audience to believe that her $\theta_D^I$ may not be small. Knowing this, D makes a strategic analysis of the game based on her preferences and beliefs over the audiences' inference of $\theta_D^I$ conditional on her actions, and acts accordingly to maximize her expected psychological utility. The nature of the interaction and the dependence of utility on information at end nodes and on endogenous beliefs make this a Psychological Game with signalling and uncertainty over private values.

Therefore, applying (2) and (5) to the Dictator Minigame, D will choose $s$ over $k$ for:

$$u_D(s, \alpha_{-D}, \theta_D, n) \geq u_D(k, \alpha_{-D}, \theta_D, n)$$

$$\pi_D(s) + \theta_D^I[1] + n\theta_D^R \mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}\left[\tilde{\theta}_D^I | s\right]\right] \geq \pi_D(k) + \theta_D^I[-1] + n\theta_D^R \mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}\left[\tilde{\theta}_D^I | k\right]\right]$$

$$\theta_D^R \geq \frac{\pi_D(k) - \pi_D(s) - 2\theta_D^I}{n\left(\mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}\left[\tilde{\theta}_D^I | s\right]\right] - \mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}\left[\tilde{\theta}_D^I | k\right]\right]\right)} \equiv \underline{\theta}_D^{R,1} \qquad 6$$

(6) implies that, all else equal, Dictators will prefer *split* over *keep* for reputational concerns high enough. Moreover, in the game, action $s$ sends ambiguous signals over D's intrinsic-motivation trait – the altruistic action may be motivated by reputational concerns rather than selfless reasons – but $k$ is an unambiguous signal of a low-intrinsic (and reputational) trait.[17] In light of this, rational players who update their beliefs according to the laws of conditional probability should expect the intrinsic trait of $i$ to be more likely to be low after observing action $k$. Ds are aware of this, and therefore the difference $\mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}\left[\tilde{\theta}_D^I | s\right]\right] - \mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}\left[\tilde{\theta}_D^I | k\right]\right]$ is larger than 0. Consequently, the denominator in (6) is increasing in $n$ and, as $n$ grows, $\underline{\theta}_D^{R,1}$ decreases linearly and the sharing condition is more likely to hold for a randomly selected D.

Assuming that $\theta_D^R$ is drawn from a continuous distribution $\psi$ with full support in $[0, \bar{\theta}_D^R]$, such a drop in $\underline{\theta}_P^{R,1}$ should in turn raise the frequency of *split* in a randomly drawn

[17] Since (2) implies that agents do not care about others' opinion over their reputational trait, the signal of a low reputational trait is irrelevant for the strategic analysis, and I will ignore it. More complete formulations of image concerns may also account for this.

population. In an experiment involving the Dictator Minigame, larger audiences should then be expected to lead to a higher frequency of *split*, but the marginal effects of increases in the audience size on the individual incentives to *split* cannot be empirically evaluated without knowledge on the distribution of traits $\psi$ and on how Ds estimate the audience inference of $\tilde{\theta}_D^I$.

If reputational concerns grow at a constant pace with the size of the audience, each spectator has the same marginal impact on D's reputational incentives to play $s$. Consequently, for a D with given reputational trait $\theta_D^R$, (6) could be rewritten to show that there exists a minimum audience size such that she will choose $s$ over $k$:

$$n \geq \frac{\pi_D(k) - \pi_D(s) - 2\theta_D^I}{\theta_D^R \left( \mathbb{E}_{\beta_D} \left[ \mathbb{E}_{\alpha_{-D}} \left[ \tilde{\theta}_D^I | s \right] \right] - \mathbb{E}_{\beta_D} \left[ \mathbb{E}_{\alpha_{-D}} \left[ \tilde{\theta}_D^I | k \right] \right] \right)} \geq \underline{n} \qquad 7$$

### 2.3.3 Root dependency: application to the Dictator Mini-Game

Reputational concerns could instead be increasing at a decreasing marginal rate in the size of the audience. The evolutionary justification for these psychological preferences may lay on gossip and the limited size of human societies during most of our evolutionary history. As information spreads within a circumscribed social network (such a pre-Neolithic tribe), multiple spectators to an act could communicate what they witness to the same person, making information redundant. If everyone communicates with a limited number of people within the community, each additional spectator increases the crowd of potential gossip recipients, but at decreasing marginal rate since every new audience member is more likely to pass on info to individuals who have already been informed by others. Put in this way, audience concerns may even still be linear in the number of people holding a specific belief but, due to gossip within a limited social group, concave in the overall number of those directly observing the act.

A simple way to model this nonlinearity is to exploit the simplifying assumptions in (4) to rewrite reputational concerns in (5) as the square root of $n$ times the reputational concern of $i$ towards any player and their common inferred intrinsic trait of $i$:

$$\sqrt{\sum_{j\neq i}\left(\theta_{ij}^{R}\mathbb{E}_{\alpha_j}\left[\tilde{\theta}_i^I|z\right]\right)^2}=\sqrt{n}\cdot\theta_i^R\mathbb{E}_{\alpha_{-i}}\left[\tilde{\theta}_i^I|z\right] \quad (8)$$

Reputational concerns in (8) are now concave in $n$. The condition to share from (5) can now be rewritten as:

$$\theta_D^R\geq\frac{\pi_D(k)-\pi_D(s)-2\theta_D^I}{\sqrt{n}\left(\mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}\left[\tilde{\theta}_D^I|s\right]\right]-\mathbb{E}_{\beta_D}\left[\mathbb{E}_{\alpha_{-D}}\left[\tilde{\theta}_D^I|k\right]\right]\right)}\equiv\underline{\theta}_D^{R,2} \quad (9)$$

The threshold $\underline{\theta}_D^{R,2}$ for sharing is still decreasing with the audience size, but at a decreasing marginal rate. In effect, for each $n>1$, it can be shown that $\underline{\theta}_D^{R,2}>\underline{\theta}_D^{R,1}$. Since $\theta_D^R$ is drawn from the same distribution, $\psi$, independently of the functional form, for all values of $n$ larger than 1, reputational concerns in (8) would predict a lower frequency of $s$ with respect to the linear version in (5). This can be shown by the fact that the cumulative distribution of the trait in the population is greater using the threshold derived from the linear function with respect to that of the root one:

$$\int_{\underline{\theta}_D^{R,1}}^{\overline{\theta}_D^R}\psi(\theta_D^R)>\int_{\underline{\theta}_D^{R,2}}^{\overline{\theta}_D^R}\psi(\theta_D^R)$$

for all values of audiences of size larger than 1, since $\underline{\theta}_D^{R,2}>\underline{\theta}_D^{R,1}$ for all $n>1$.

# 3 Experimental Design

## 3.1 Experimental procedures and treatments

This experiment is designed to be run on computers in a laboratory setting, ideally the Bocconi Experimental Laboratory for Social Sciences (BELSS) at Bocconi University. Subjects will be drawn randomly among the University student body through email invitations, on-campus posters, and/or using BELSS' Sona-System for recruitment.

Participants will be randomly assigned to a role in the Dictator Minigame with external audience – Figure 2 – The Dictator Minigame with external audience. The three roles are: Dictator (hereinafter called the "Proposer", or "P", for neutrality), Receiver, and Audience member. Proposers are the only active players in the game, while Receivers and Audience members are passive observers who wait for the Proposer to decide the distribution of the endowment and can perfectly observe outcomes. Roles are anonymously matched, and each player only knows her role and the number of co-players who can observe the interaction on their screens (including the audience). Players do not know the identity of their co-players but know that they are matched with other people in the lab. While anonymity may significantly reduce the magnitude of reputational concerns, it consolidates internal validity by sidestepping confounders related with privacy, identifiability, face-to-face interactions, and expectations of potential interactions outside the laboratory. Moreover, if evidence for size-dependent audience effects can be gathered under anonymous interactions, subsequent experiments inducing some degree of identifiability could follow to further analyse the strength of the interaction between reputational concerns and the number of spectators.

Participants in the experiment are randomly assigned to treatments, which differ in the number of audience members observing the interaction. Treatment Base involves zero external observers, Treatment 1 has one, Treatment 2 has three, and Treatment 3 has five. Instructions for the subjects are unchanged across treatments. Table 2 summarizes the different treatments.

*Table 2*

| | TB | T1 | T2 | T3 |
|---|---|---|---|---|
| Number of external spectators | 0 | 1 | 3 | 5 |

At the beginning of the experimental session, participants are randomly assigned to a computer seat and each computer will open to a page which assigns every participants to a role. All forms of communication are strictly forbidden during all experimental sessions.

At the beginning of the game, each Proposer receives an endowment of 10 Experimental Currency Units, or ECU (1 ECU = €1). The Proposer has a dichotomous choice between two options labelled as *split* or *keep. Split* divides the endowment equally between P and R (resulting in material outcomes of 5 ECU for each player); *keep* holds most of the endowment in P's pockets (yielding payoffs 8 ECU for P and 2 ECU for R). The Audience receives a fixed payoff of 2 ECU which is unaffected by the outcomes of the games observed by them. The presence and size of the external audience is determined by the treatment assigned to the couple of coplayers. On top of the payoffs from playing the game, each player also receives a show-up fee of 3 ECU. The smallest amount a participant can expect to receive from taking part in the experiment is hence 5 ECU, the maximum is 11. Each experimental session is expected to last around 20 minutes in total, including explanations, roles allocation and a final questionnaire on demographics. This is in line with a wage of €15 per hour or more.

The rules of the game and the audience size are common knowledge to everyone. The number of observers will be displayed on everyone's screen as the Proposer makes her choices. At the end of the game there is perfect terminal information about P's action and all players are informed about the material outcomes of their coplayers. P's action is communicated clearly at the end of the game to ensure an unmistakable link between her action and the outcomes realized, hence ensuring reputational accountability.

For the actual experiment, four sessions involving one treatment each and with subjects playing only one game will be carried out autonomously and possibly on the

same day. The desired number of observations will be determined based on a pilot study. The pilot will evaluate all treatments in a single session and involve a total of 12 couples of active players plus observers. Each non-audience subject will play twice in the pilot, once as the Dictator and once as the Receiver, with a different and anonymous coplayers every time. Since this could rise concerns related to various confounders, the actual experiment requires that each player only plays once. The pilot will produce 6 observations per treatment for a total of 24 observations. Its results will be used to compute, through power analysis, the adequate number of observations and sessions required to properly assess the experimental evidence.

As mentioned, other-regarding preferences influence P's decision (for experimental evidence, see Camerer and Thaler (1995); Charness and Dufwenberg (2006); Battigalli and Dufwenberg (2007); Bellemare et al. (2017); Attanasi et al. (2019)), but should be unchanged in the number of external observers in the different treatments. Significant differences in the frequency of the action *split* between treatments can then be attributed to changes in the magnitude of reputational concerns.

## 3.2 Assumptions, hypotheses, and behavioural predictions

The aim of the experiment is to assess whether and how Proposers' behavior in the Dictator Minigame may be subject to *audience effect*s. The following assumptions and related experimental hypotheses summarize the theory introduced and its behavioural predictions in the Dictator Minigame.

Assumption 1:

*Most people have image concerns related to being perceived as altruistic.*

From Assumption 1, I derive Hypothesis 1:

*The presence of an external audience increases the frequency of split.*

Therefore, I expect TB, where no external audience is present, to have a significantly lower frequency of *split* (altruistic) allocations with respect to T1, T2 and T3:

$$\mathrm{Fr}_{\mathrm{split}}(\mathrm{TB}) < \mathrm{Fr}_{\mathrm{split}}(\mathrm{T1} \wedge \mathrm{T2} \wedge \mathrm{T3}) \qquad 10$$

Assumption 2:

*The larger the audience, the greater the magnitude of image concerns.*

From Assumption 2, Hypothesis 2:

*The larger the number of audience members, the higher the frequency of split.*

Therefore, I expect the frequency of *split* to be increasing in the number *n* of audience members witnessing the play:

$$\mathrm{Fr}_{\mathrm{split}}(\mathrm{T3}) > \mathrm{Fr}_{\mathrm{split}}(\mathrm{T2}) > \mathrm{Fr}_{\mathrm{split}}(\mathrm{T1}) > \mathrm{Fr}_{\mathrm{split}}(\mathrm{TB}) \qquad 11$$

Assumption 3:

*Image concerns are increasing at decreasing marginal rate in the number of spectators.*

From Assumption 3, Hypothesis 3 follows:

*Each additional Audience Member has a smaller impact on the frequency of split*

Therefore, the relationship between the audience and the frequency of *split* should be a concave function of the number of Audience Members.

## 3.3 Analysis of the data

The hypotheses will be tested through a simple linear probability model estimated through a least squares regression analysis using a linear or quadratic fit. The dependent variable of interest is the probability of the Proposer playing *split*. The baseline is the probability of playing *split* with the Receiver as the only observer and the independent variable is the number of external observers witnessing the play. If the coefficient of the independent variable is positive and significant, the hypothesis 2 (11) is verified.

To study whether an external audience, of any size, impacts on the likelihood of playing *split*, all treatments but TB could be clustered under a dummy variable, $S$ for Spectators. If the coefficient of $S$ is positive and statistically significant it will serve as evidence that external audiences in general can increase the likelihood of prosocial behaviour, confirming hypothesis 2 (10).

To test hypothesis 3 instead, I will study whether a quadratic regression fits the data better than a linear one. Yet, without any knowledge about the individual preferences and beliefs of Proposers and how those are distributed among the population, the marginal impact of an increase in number of spectators on the frequency of action

*split* can only be evaluated in aggregate, with limited scope for causal considerations about how the audience size impacts individual reputational concerns at the margins.

## 3.4 Discussion and Extensions

### 3.4.1 Belief elicitation I: guilt and size-dependent expectations

Up to this point, I set aside the potential confounding impact of guilt (and additional belief-dependent, other-regarding preferences) on the ground that those are unchanged in the size of the audience. Yet, awareness of the presence of spectators observing P's action may increase the Receivers' expectations, meaning that they may, intuitively or rationally, expect Ps to be more likely to *split* under larger audiences. If this is true, larger audiences may lead to higher Receivers' expectation, disappointment from action *keep*, and thus higher guilt for Proposers (Battigalli and Dufwenberg (2007)). In turn, this would further incentivize guilt-averse players to play *split*, with a multiplying effect of the audience size on the frequency of *split* which is mediated by more than reputational concerns.

One way to circumvent this problem is to ensure that Rs' expectations are constant in the size of the audience, and that Ps are aware of this. This could be done by slightly modifying the design of the game by communicating the size of the external audience only to the Proposer. Yet, size-dependent expectations of the frequency of action *split* may be further evidence in support of auxiliary beliefs supporting audience effects and thus be worthy of record.

Those expectations could be collected and compared across treatments by asking everyone (Ps, Rs, and Audience Members) out of 10 other groups in the laboratory playing with their same audience size, how many will play *split*.[18] Monetary incentives for correct guesses can be used to ensure higher reliability of the data. Moreover, the information collected could also be useful to study if people intuitively gauge the effect of audiences and, in the presence of larger audiences, expect others to be more altruistic.

[18] Belief elicitation should not affect the results, as it should not have any significant impact on subject's behaviour, see Bauer and Wolff (2018).

### 3.4.2 Belief elicitation II: compliance to a social norm

The model developed in this work was grounded on the further simplifying assumption (3) that only 'good' reputations, coming from altruistic actions, have an impact on the psychological utility of players. Alternative models of reputational concerns involve compliance to a social norm (e.g., Bernheim (1994)). In those models, reputation-motivations depend on what players believe that others expect them to do. Applied to the Dictator Minigame, Receiver's motivations would then depend on what they believe that others think the social norm is. While under a social norm for altruism this class of models would lead to the same predictions already made, if Proposers believe that they are expected to behave selfishly, reputational concerns may induce them to play *keep*.

This could be the case, for instance, for Economics students with some knowledge of traditional Game Theory. In an academic environment, they could expect spectators to be familiar with the Nash-dominant strategy of the game form and believe that observers may expect them to play it. A failure to play the dominant strategy may be seen as a deviation from some norm (playing the dominant strategy) or become an undesirable signal of poor academic knowledge (not knowing what the dominant strategy is).

Setting Receivers' aside – whose expectation may be skewed by wishes and who may be relevant for other-regarding preference – eliciting Ps' beliefs about the external audience's expectations can help determine if their behaviour is influenced by some social norm. If the frequency of *split* is correlated with Ps' elicited beliefs over what they think others expect them to play, for all possible beliefs about the audience's expectation, then players are more concerned about complying with a norm then about displaying their selflessness. An additional Assumption, related Hypothesis, and survey question could then be introduced to verify compliance to a social norm, proxied by others' expectations.

Assumption 4:

*People's actions are influenced by beliefs over what observers expect them to do.*

From Assumption 4, Hypothesis 4 follows for the empirical analysis:

*Proposers' decision to share correlates positively with their belief over what the audience expects them to play.*

Beliefs about the social norm can be elicited by asking Proposers in treatments 1 to 3 how many, out of 10 Audience Members, they believe expect Proposers to play *keep*. If Ps' actions correlate with their beliefs – both when P believes that there exists a social norm towards playing *keep* and when she believes that the norm is for altruism – then models relying on social norms may predict behaviour better. Such a result would also support evolutionary theories of group-level selection by showing how social groups and cultures partly shaped the human mind for groupishness and large-scale cooperation through conformity and norms compliance.

### 3.4.3 Subjects' morality, preferences, and characteristics

To further ensure that the results are exclusively driven by reputational concerns, the empirical analysis should be run while controlling for a measure of the subjects' morality. The Aquino questionnaire for measuring the self-assessed importance of moral values could serve this goal. To avoid nudging subjects' behaviour towards more 'moral' actions (if they declared a high level of morality), the survey should be filled after playing the Game.

Subjects could also differ in their reputational preferences. For instance, demographic factors could shape the sort of reputation that they wish to attain. While the simplifying assumption in (3) ruled this out for this study, subjects may wish to signal their toughness over altruism, and size-dependent reputational concerns may run in opposite to the hypotheses made. One way to control for players' heterogeneity of reputational preferences is to insert, among the general demographic questions in the questionnaire, whether they would rather appear as "tough" or "altruistic". However, it is unclear to which extent declaring a certain preference could influence behaviour (or the other way around), nudging results towards a correlation that may not be there otherwise. A fifth assumption, and related hypothesis could be investigated to verify the presence of this confounder.

Assumption 5:

*Audience effects are aligned with people's reputational preferences*.

From A5, Prediction 5:

*Audience effects increase the frequency of split only among Proposers who care about being perceived as altruistic*.

Therefore, Proposers reporting a higher level of concern for being perceived as altruistic (with high intrinsic motivations) should show a higher frequency of split than those who care about being perceived as tough (not pushovers or more selfish).

Audience demographic and social characteristics could also be a factor. Subjects may wish to attain different reputations with different audiences (as in Attanasi et al. (2019)) or feel more inclined to undertake altruistic actions towards similar people (Haidt (2012)). It could be interesting, for instance, to understand to what extent people's altruism is shaped by parochiality: audiences' characteristics could be selectively revealed in subsequent studies to see if audience effects are enhanced by perceived similarities with the audience. Features such as belonging to the same moral community (e.g., sharing religious faith or political orientation) or having common social or ethnic characteristics (such as a going to the same university or coming from the same racial background) may all be significant factors in shaping parochial altruism (Haidt (2012)). As argued by Haidt (2012), if selflessness shot up due to group-level selection pressures it should be more pronounced towards those perceived as fellow members of a social (and moral) community. Audience features other than size may then be investigated to learn how to incentivize altruistic behaviour.

Sex (or gender) could also be a relevant factor shaping audience effects. An audience entirely comprised of potential mating competitors could trigger different drives than one made up of potential mates. Accordingly, the sex (or gender) of Receivers could also influence behaviour, as well as interact with that of the audience to signal different traits (such as, for a male Proposer, signalling to an audience of females his ability to cooperate with a male Receiver). Culture is also likely to mediate those effects.

### 3.4.5 Other possible confounders

While simple and straightforward, the experimental design has some other limits, which are mostly related to striking a balance between different motivations.

Firstly, in the instruction of the experiment, the action label "*keep*" induces endowment effects, making Proposers more resistant to conceding a share of what they consider part of their own endowment (Kahneman (2011)). While this is purposely done to study altruism and how it is impacted by audiences, if the magnitude of the effect is substantial I could be left with too few observations of *split*.

The payoffs of the experiments are also designed to avoid extreme allocations: Proposers cannot leave Receivers with less than 2 ECU (plus the show-up fee). This is done to tame the selfishness of action *keep*, preventing it from being too brutal. Yet, if R's payoff in *keep* is too generous the experiment could once again produce too few observations for action *split*.

Moreover, Rs' overall payment under *keep* is equivalent to the Audience's, and this is done to reduce the incentives to play *split* coming from others-regarding inequity aversion. In fact, if the Audience earned more than R, P may want to play *split* to reduce the inequity between R and the Audience, and this motivation could also be increasing depending on the size of the Audience, challenging the results.

Furthermore, reducing P's action to a dichotomous choice between a selfish and a generous action may diminish the richness of the data collected, as minor variations in reputational motivations may fail to induce behavioural changes. Despite this, reducing the choice to only two options allows me to unequivocally assert one choice as selfish and the other as fair, cutting down potential confounders coming from individual interpretations of the payoffs.

But the major weakness of the experiment comes perhaps from the poor representativeness of the population from which the subjects are drawn, which impacts negatively on the external validity. In an extensive review of data coming from a wide range of studies in top behavioural science journals, Henrich et al. (2010) find that the samples employed in most experimental studies are drawn from an unrepresentative group for the human population at large. Most experiments are run

by university research centres on students, and the "standard subject" is often a "WEIRD" outlier of humankind – Western, Educated, Industrialized, Rich, and Democratic. For Henrich and his colleagues, WEIRD subjects differ on a range of domains including fairness, cooperation, moral reasoning, and even heritable IQ measures (Henrich et al. (2010)). Consequently, the results obtained through the proposed experiment should be verified for a more varied and representative sample of the global population, at least for what concerns the level of education and age.

Other common confounders and issues related to laboratory experiments in Economics include the Hawthorne effect and deception. The Hawthorne effect is the well-studied tendency to behave differently under observation (Landsberger (1957)). Despite the presence of economic incentives, subjects aware of being under observation from the experimenter may alter their behaviour, especially in a laboratory setting, challenging the potency of the causal evidence in the findings.

Deception is instead only related to a more practical issue of the experimental design: its prohibitive costs. In order to produce one observation, the actual experiment would require between two (in TB) and seven (in T3) subjects. This implies costs ranging between €16 and €31 per observation, according to the treatment. To circumvent this problem, a single audience could be used for all groups in one experimental session. Each proposer could only be informed of the presence of a certain number of spectators observing her action, without specifying how many other interactions her observers are witnessing. Yet, due to extremely strict anti-deception norms in Economics, this may amount to deception for some Journals, making the study unpublishable. The cost-savings of the modification for a pilot of 24 interactions, equally spread across the 4 treatments, with every subject (including the audience) playing twice, would amount to €154 – the totals are €311 with a single audience for the whole session, compared to €465 with an exclusive audience for each interaction.

# 4 Conclusions

This paper aimed at discussing why and how audience effects could impact agents' behaviour in strategic interactions via their reputational concerns. The "why" question has been investigated through a broad review of the psychological and evolutionary origins of altruism and audience concerns. The "how" question has been explored by studying a theoretical model to account for size-dependent audience effects and by proposing a theory-driven experimental design to test its implications.

After having presented the traditional approach to altruism in Economics, some evolutionary explanations for selflessness were presented and grouped according to the main innate behavioural drive supporting them: kinship, reciprocity, or groupishness. Notably, reputation-driven audience effects could be justified by natural selection on both individual grounds – via reciprocity and gains from cooperation – and collective grounds – via group-level advantages of norm enforcement. The empirical evidence from studies in economics and psychology analysed in this paper seems consistent with a mix of the two. In turn, the size of an observing audience should affect the magnitude of reputational concerns.

The main theoretical contribution of the paper amounts to adapting an existing model of reputational concerns, borrowed from Battigalli and Dufwenberg (2020), to the Dictator Minigame to investigate the impact of variations in the audience size on psychological incentives. Two game-specific functions for reputational concerns were then developed under different assumptions about the marginal impact of additional observers, and I proposed a related experimental design to test the predictions of the theory.

Finally, I discussed extensively the limits of the theoretical model and of the related experiment, highlighting the constraints in the design and proposing possible extensions to account for a wider range of psychologically plausible reputational motivations.

# References


ABELER, J., D. NOSENZO, AND C. RAYMOND. (2019): "Preferences for Truth-Telling," *Econometrica*, 87, 1115–53.

ALEXANDER, R. D. (1987): "The Biology of Moral Systems," New York, *Transaction Publishers*.

ANDREONI, J., AND D. B. BERNHEIM. (2009): "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, 77, 1607–36.

ANDREONI, J. (1989): "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy*, 97, 1447–58.

AQUINO, K., AND A. REED. (2002): "The self-importance of moral identity.," *Journal of Personality and Social Psychology*, 83, 1423–40.

ARIELY, D., A. BRACHA, AND S. MEIER. (2009): "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *The American Economic Review*, 99, 544–55.

ATTANASI, G., R. DESSÍ, F. MOISAN, D. ROBERTSON, D. ARIELY, S. CHOI, F. ETILÉ, ET AL. (2019): "Public Goods and Future Audiences: Acting as Role Models?," *Social Macroeconomics Working Paper Series*.

BATTIGALLI, P., AND M. DUFWENBERG. (2007): "Guilt in Games," *The American Economic Review*, 97, 170–76.

---. (2009): "Dynamic psychological games," *Journal of Economic Theory*, 144, 1–35.

---. (2020): "Belief-Dependent Motivations and Psychological Game Theory," *Journal of Economic Literature*, forthcoming.

BATTIGALLI, P., R. CORRAO, AND M. DUFWENBERG. (2019): "Incorporating belief-dependent motivation in games," *Journal of Economic Behavior & Organization*, 167.

BAUER, D., AND I. WOLFF. (2018): "Biases in Beliefs: Experimental Evidence," *TWI Research Paper Series*, 109.

BELLEMARE, C., A. SEBALD, AND S. SUETENS. (2017): "A note on testing guilt aversion," *Games and Economic Behavior*, 102, 233–39.

BÉNABOU, R., AND J. TIROLE. (2006): "Incentives and Prosocial Behavior," *The American Economic Review*, 96, 1652–78.

BERNHARD, H., U. FISCHBACHER, AND E. FEHR. (2006): "Parochial altruism in humans," *Nature*, 442, 912–15.

BERNHEIM, D. B. (1994): "A Theory of Conformity," *Journal of Political Economy*, 102, 841–77.

BÖHEIM, R., D. GRÜBL, AND M. LACKNER. (2019): "Choking under pressure – Evidence of the causal effect of audience size on performance," *Journal of Economic Behavior & Organization*, 168.

BUSS, D. M. (1998): "Evolutionary Psychology: The New Science of the Mind," New York, *Routledge*.

CAMERER, C., AND R. H. THALER. (1995): "Anomalies: Ultimatums, Dictators and Manners," *Journal of Economic Perspectives*, 9, 209–19.

CAÑIGUERAL, R., AND A. F. DE C. HAMILTON. (2019): "Being watched: Effects of an audience on eye gaze and prosocial behaviour," *Acta Psychologica*, 195, 50–63.

CARNEIRO, R. L. (1970): "A Theory of the Origin of the State: Traditional theories of state origins are considered and rejected in favor of a new ecological hypothesis," *Science*, 169, 733–38.

CHARNESS, G., AND M. DUFWENBERG. (2006): "Promises and Partnership," *Econometrica*, 74, 1579–1601.

COTTRELL, N. B., D. WACK, G. J. SEKERAK, AND R. H. RITTLE. (1968): "Social Facilitation of Dominant Responses by the Presence of an Audience and the Mere Presence of Others," *American Psychological Association.*

COSMIDES, L. (1985): "Deduction or Darwinian algorithms? An explanation of the "elusive" content effect on the Wason selection task". *Harvard University*.

DARWIN, C. (1859): "On the Origins of the Species," London, *John Murray*.

---. (1871): "The Descent of Man, and Selection in Relation to Sex," London, *John Murray*.

---. (1872): "The Expression of the Emotions in Men and Animals," London, *John Murray*.

DAWKINS, R. (1976): "The Selfish Gene," Oxford, *Oxford University Press*.

DE CREMER, D., AND M. BAKKER. (2003): "Accountability and cooperation in social dilemmas: The influence of others' reputational concerns," *Current Psychology*, 22, 155–63.

DELLAVIGNA, S., J. A. LIST, AND U. MALMENDIER. (2012): "Testing for Altruism and Social Pressure in Charitable Giving," *The Quarterly Journal of Economics*, 127, 1–56.

DELTON, A. W., M. M. KRASNOW, L. COSMIDES, AND J. TOOBY. (2011): "Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters," *Proceedings of the National Academy of Sciences*, 108, 1335–40.

DUFWENBERG, M., AND G. KIRCHSTEIGER. (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47, 268–98.

EMBER R., EMBER, M., MARTIN, D. L., AND FRAYER, D. W. (1997): "Violence in the ethnographic record: Results of cross-cultural research on war and aggression," in *Troubled times: Violence and warfare in the past*, *3*, 1-20, Amsterdam, *Gordon and Breach*.

FEHR, E., AND B. ROCKENBACH. (2004): "Human altruism: economic, neural, and evolutionary perspectives," *Current Opinion in Neurobiology*, 14, 784–90.

FEHR, E., AND K. M. SCHMIDT. (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114, 817–68.

FEHR, E., AND FISCHBACHER, U. (2003): "The nature of human altruism," *Nature*, 425, 785–91.

FISCHBACHER, U., AND F. FÖLLMI-HEUSI. (2013): "Lies in disguise an experimental study on cheating," *Journal of the European Economic Association*, 11, 525–47.

GAZZANIGA, M. S., R. B. IVRY, AND G. R. MANGUN. (2014): "Ch. 10, Emotion," in *Cognitive Neurosciences*, WW Norton & Co.

GEANAKOPLOS, J., D. PEARCE, AND E. STACCHETTI. (1989): "Psychological games and sequential rationality," *Games and Economic Behavior*, 1, 60–79.

GLOWACKI, L., M. L. WILSON, AND R. W. WRANGHAM. (2017): "The evolutionary anthropology of war," *Journal of Economic Behavior & Organization*, 178.

GNEEZY, U., A. KAJACKAITE, AND J. SOBEL. (2018): "Lying Aversion and the Size of the Lie," *American Economic Review*, 108, 419–53.

HAIDT, J. (2012): "The righteous mind: Why good people are divided by politics and religion". London, *Penguin*.

HAMILTON, A. F. DE C., AND F. LIND. (2016): "Audience effects: what can they tell us about social neuroscience, theory of mind and autism?," *Culture and Brain*, 4, 159–77.

HAMILTON, W. D. (1964a): "The genetical evolution of social behaviour. i," *Journal of Theoretical Biology*, 7, 1–16.

---. (1964b): "The genetical evolution of social behaviour. ii," *Journal of Theoretical Biology*, 7, 17–52.

HARE, B., V. WOBBER, AND R. WRANGHAM. (2012): "The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression," *Animal Behaviour*, 83, 573–85.

HENRICH, J., S. J. HEINE, AND A. NORENZAYAN. (2010): "The weirdest people in the world?," *Behavioral and Brain Sciences*, 33, 61–83.

JAMES, W. (1884): "What is an Emotion?," *Mind*, 9, 188–205.

JAMES, W, DRUMMOND, R. (1890): "The principles of psychology".

KAHNEMAN, D. (2011): "Thinking, fast and slow". London, *Macmillan*.

KAHNEMAN, D., AND A. TVERSKY. (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263.

KEELEY, L. H. (1996): "War before civilization". New York, *Oxford University Press*.

KHALMETSKI, K., AND D. SLIWKA. (2019): "Disguising Lies—Image Concerns and Partial Lying in Cheating Games," *American Economic Journal: Microeconomics*, 11, 79–110.

KRASNOW, M. M., L. COSMIDES, E. J. PEDERSEN, AND J. TOOBY. (2012): "What Are Punishment and Reputation for?," *Plos One*, 7, e45662.

LANDSBERGER, H. A. (1957): "Hawthorne Revisited: A Plea for an Open City," Ithaca, N.Y.: Cornell University, *Open WorldCat*.

LEARY, M. R., AND R. F. BAUMEISTER. (2000): "The nature and function of self-esteem: Sociometer theory," *Advances in Experimental Social Psychology Volume 32*, 32, 1–62.

LINARDI, S., AND M. MCCONNELL. (2008): "Volunteering and Image Concerns," *undefined*.

NEUMANN, J. VON, AND O. OSKAR MORGENSTERN. (1944): "Theory of Games and Economic Behavior," New Jersey, *Princeton University Press*.

OZBAY, E. F., AND E. Y. OZBAY. (2013): "Effect of an audience in public goods provision," *Experimental Economics*, 17, 200–214.

PETERSEN, M. B., D. SZNYCER, L. COSMIDES, AND J. TOOBY. (2012): "Who Deserves Help? Evolutionary Psychology, Social Emotions, and Public Opinion about Welfare," *Political Psychology*, 33, 395–418.

PICKRELL, J. K., G. COOP, J. NOVEMBRE, S. KUDARAVALLI, J. Z. LI, D. ABSHER, B. S. SRINIVASAN, ET AL. (2009): "Signals of recent positive selection in a worldwide sample of human populations," *Genome Research*, 19, 826–37.

PINKER, S. (1997): "How the Mind Works," New York, *WW Norton & Co*.

RABIN, M. (1993): "Incorporating Fairness into Game Theory and Economics," *The American Economic Review*, 83, 1281–1302.

SEBALD, A., AND N. VIKANDER. (2018): "Optimal firm behavior with consumer social image concerns and asymmetric information," *Journal of Economic Behavior & Organization*, 167.

SYMONS, D. (1992): "On the use and misuse of Darwinism in the study of human behavior", In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), "The adapted mind: Evolutionary psychology and the generation of culture" (p. 137–159). *Oxford University Press*.

SOETEVENT, A. R. (2011): "Payment Choice, Image Motivation and Contributions to Charity: Evidence from a Field Experiment," *American Economic Journal: Economic Policy*, 3, 180–205.

THALER, R. H. (1988): "Anomalies: The Ultimatum Game," *Journal of Economic Perspectives*, 2, 195–206.

TOMASELLO, M., M. CARPENTER, J. CALL, T. BEHNE, AND H. MOLL. (2005): "Understanding and sharing intentions: The origins of cultural cognition," *Behavioral and Brain Sciences*, 28, 675–91.

TOOBY, J., AND L. COSMIDES. (2010): "Groups in Mind: The Coalitional Roots of War and Morality," in *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, ed. by Boesch, C. and L. Cosmides, *Macmillan*, 91–234.

---. (2005). "Conceptual Foundations of Evolutionary Psychology," In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (p. 5–67). Santa barbar, *John Wiley & Sons*.

TRIPLETT, N. (1898): "The Dynamogenic Factors in Pacemaking and Competition," *The American Journal of Psychology*, 9, 507.

TRIVERS, R. L. (1971): "The Evolution of Reciprocal Altruism," *The Quarterly Review of Biology*, 46, 35–57.

WILSON, D. S. (1992): "On the relationship between evolutionary and psychological definitions of altruism and selfishness," *Biology & Philosophy*, 7, 61–68.

ZAJONC, R. B. (1965): "Social Facilitation," *Science*, 149, 269–74.

# Appendix

## General instructions for the subjects

Hello, welcome to this experiment.

Before starting, we kindly remind you that all forms of communication are strictly forbidden during the experiment.

The experiment will be carried out as follows: you will be assigned to a role. The experiment will consist in only two interactions, in which you will change role and coplayers. The three possible roles are either *Proposer*, *Receiver,* or *Audience Member* (a passive spectator whose payment is unaffected by the game).

If you are assigned to the role of *Proposer*, you can take one action that affect your payment and that of the *Receiver* player.

*Proposers* and *Receivers* will be matched in random pairs and will play a Proposer game, in which each *Proposer* has 2 different options to divide an endowment of 10 Euro:

- option 1 – *split*: Proposer and Receiver both get 5 Euro.
- option 2 – *keep*: the Proposer gets 8 Euro and the Receiver gets 2 Euro.

The game ends after the Proposer has decided how to split the endowment. Receivers are passive, they will just observe the outcome of their game and earn their payoff.

Audience members (if present) will receive a fixed payoff of 2 Euro for their role, which is unaffected by the outcome of the game. Everyone is informed of the presence (or absence) and size of the audience. Please remember that your identity as well as that of other players will remain anonymous.

On top of the payoff from the game, every participant will receive a 3 Euro show up fee for coming here today. The minimum payoff one can receive is therefore 7 Euro (3 Euro plus 4 Euro minimum from playing two games).

Thank you for your participation!

## Instructions for the game

### Proposers in TB

You have been assigned the role of Proposer.

You will decide how to split 10 Euro between you and the Receiver according to 2 options:

- Option *split*: both you and the Receiver get 5 Euro.
- Option *keep*: you keep 8 Euro, Receiver gets 2 Euro.

Note that the Audience's payoff of 2 Euro is unaffected by your choice.

No one is observing your choice other than the Receiver.

Please, click on the box with the option that you want to choose

| Split | Keep |
|---|---|

### Receivers in TB

You have been assigned to the role of Receiver.

The Proposer will decide your share of an endowment of 10 Euro according to his or her 2 available options:

- Option *split*: both you and the Proposer get 5 Euro.
- Option *keep*: the Proposer gets 8 Euro, you get 2 Euro.

Note that the Audience's payoff of 2 Euro is unaffected by this choice.

No one is observing the Proposer choice and your payoffs other than you.

### Proposers in T1, T2, T3

You have been assigned the role of Proposer.

You will decide how to split 10 Euro between you and the Receiver according to 2 options:

- Option *split*: both you and the Receiver get 5 Euro.
- Option *keep*: you keep 8 Euro, Receiver gets 2 Euro.

Note that the Audience's payoff of 2 Euro is unaffected by your choice.

Please, notice that __ people are observing your game and your choice.

Please, click on the box with the option that you want to choose

| Split | Keep |
|---|---|

### Receivers in T1, T2, T3

You have been assigned to the role of Receiver.

The Proposer will decide your share of an endowment of 10 Euro according to his or her 2 available options:

- Option *split*: both you and the Proposer get 5 Euro.
- Option *keep*: the Proposer gets 8 Euro, you get 2 Euro.

Note that the Audience's payoff of 2 Euro is unaffected by this choice.

Please, notice that __ people are observing the Proposer's choice and resulting payoffs.

## Beliefs elicitation

Before eliciting the subject's beliefs, an economic incentive to be determined shall be communicated and paid for correct (or very close) answers.

### Receivers' initial expectations in TB to T3

Out of 10 couples playing the Proposer game, with your same number of spectators, in how many do you think the Proposers will play *split*? _____

### Audience's initial expectation in treatments T1 to T3

Out of 10 couples playing the Proposers game, in how many do you think the Proposers will play *split*? _____

### Proposers' initial beliefs

Out of 10 Audience Members observing someone playing the Proposers game, how many Audience Members do you think expect the Proposers to play *split*? _____

## Proposers' reputational preferences

Do you care more about being perceived by others as *altruistic* or *tough*?

| Altruistic | Tough |
|---|---|

## Demographic data and survey

Thanks for participating to the experiment! Before you go, please take a minute to answer some quick questions on basic demographics. All data will be kept anonymous and collected and treated according to GDPR regulation.

Please, enter your age: _____

Please, enter your gender (male/female/other): _____

Are you a married? (yes/no): _____

Please, enter the number of years of education completed: _____

Please, specify your ethnicity: _____

Where do you live? (Please, type city of residence): _____

Are you a student? (yes/no): _____

If yes, please, enter your field of study: _____

Are you familiar with the concept of "Nash Equilibrium"? (yes/no): _____

Are you currently employed/working? (yes/no): _____