CENTER FOR
ECONOMIC
BEHAVIOR &
INEQUALITY

CEBI WORKING PAPER SERIES

Working Paper 26/20

WHEN ARE GROUPS LESS MORAL THAN INDIVIDUALS?

Pol Campos-Mercade

ISSN 2596-44TX

CEBI
Department of Economics
University of Copenhagen
www.cebi.ku.dk

When are groups less moral than individuals?*

Pol Campos-Mercade†

September 14, 2020

Abstract

People are less likely to make moral decisions when they are in groups. I study when this phenomenon makes *groups* less likely to produce a morally desirable outcome than one individual alone. I formulate and test a model in which a moral outcome occurs if at least one individual makes a costly decision. Using a lab experiment and data from field experiments on the bystander effect, I show that if most individuals are moral, the moral outcome is more likely to be produced by one individual, whereas if most individuals are immoral, it is more likely to be produced by a group. This rule is not only useful for reconciling previous mixed evidence on moral decisions in groups, but may also be applied to better design organizations and institutions.

Keywords: Moral behavior, Group size, Bystander effect, Social preferences.

JEL Codes: C92, D64, D90.

*I would like to acknowledge the Torsten Söderbergs Stiftelse and the Royal Swedish Academy of Sciences for financial support. I am grateful to Peter Fischer for generously sharing his and his co-authors' meta-analysis data on the Bystander Effect. I thank Jordi Brandts, David Cooper, Peter Jochumzen, Nagore Iriberri, Daniele Nosenzo, Armando Meier, Erik Mohlin, Natalia Montinari, Hampus Poppius, Matthew Rabin, Pedro Rey-Biel, Alexandros Rigos, Florian Schneider, Roel van Veldhuizen, Erik Wengström, Leeat Yariv, and Jessica Zoller for their comments and suggestions. I thank participants at the CBS Behavioral and Organizational Design Workshop in Copenhagen, the European ESA Conference in Vienna, the 12th Nordic Conference in BEE in Gothenburg, the World ESA Conference in Berlin, the Microeconomics seminar in Lund, and the North American ESA conference in Antigua for their helpful comments. The experiment was pre-registered in the AEA RCT Registry (AEARCTR-0002982). Earlier versions of the paper had the titles "Helping Behavior and Group Size" and "The Group Bystander Effect."

†Department of Economics, University of Copenhagen. *e-mail*: pcm@econ.ku.dk.

1 Introduction

Many morally desirable outcomes occur if at least one person makes a moral decision. Examples of such decisions are whistleblowing on bad practice in a company, reporting a case of harassment in the workplace, volunteering for a committee, or helping a person in need on the street. In these situations, a large literature in economics and psychology shows that people are less likely to make the moral decision if they are in a group than if they are alone.[1] Darley and Latané (1968) called this phenomenon the *bystander effect* (BE).

While most research has focused on studying whether people are less likely to make moral decisions when they are in a group, policy-makers and organizations are generally more interested in the outcome rather than the individual behavior. For example, a legislator is not interested in knowing that workers in groups are less likely to whistleblow, but cares instead about whether a group of witnesses is more or less likely to report malpractice than one witness alone. Similarly, an employer is not concerned that employees in groups are less likely to help others, but rather cares about whether employees in need will be helped or not. When should one expect moral outcomes to be more likely to be produced by one individual alone rather than by a group? Despite its practical relevance, this question has not yet been the subject of any systematic study.

This paper uses a game-theoretic model, a lab experiment, and field data from the meta-analysis by Fischer et al. (2011) to show that the answer to this question critically depends on the distribution of moral preferences in the population. In situations in which most people are *moral*, in the sense that they value the moral outcome higher than the cost of making the moral decision, the moral outcome is more likely to occur when there is only one individual than when there is a group of people. The reason is that, while the moral outcome is very likely to be produced by one individual alone, it may not be produced by a group since individuals free-ride on each other. However, in situations in which few people are moral, the moral outcome is more likely to occur when there is a group than when there is only one individual. The reason is that the probability that there is at least one moral individual who

[1] See e.g. Bartling and Özdemir (2017), Behnk et al. (2017), Bergstrom et al. (2019), Campos-Mercade (2018b), Dana et al. (2007), Falk and Szech (2013), Falk et al. (2020), Fromell et al. (2017), and Panchanathan et al. (2013) for papers in economics. See also Fischer et al. (2011) for a review of the psychology literature.

is willing to make the decision is higher in larger groups. Since individuals know that most people in this situation are not moral, moral individuals take the responsibility to make the decision even when they are in a group. I show that these results can reconcile previous mixed findings in Fischer et al. (2011), where 32 (40%) studies find that people in need are more likely to be helped by one individual alone than by a group, and 48 (60%) studies find the opposite.

In the model, the key mechanism behind the results is that people in groups with a higher proportion of moral individuals are less likely to make the moral decision. This is because, if the proportion of moral individuals is high, there are more individuals with whom to share the responsibility of making the moral decision. A key challenge of testing this hypothesis is that moral preferences are typically unknown. Moreover, situations in which most people are willing to make a moral decision could systematically differ from situations in which most people are not. I circumvent these issues by conducting a pre-registered laboratory experiment in which I observe (a proxy of) the moral type of each subject and then exogenously vary the proportion of moral subjects in each game.

In the experiment, one recipient starts with €0 and is assigned either to one, two, or three dictators who start with €10 each (the 1-Dictator, 2-Dictator, and 3-Dictator games). Each dictator can then independently make the costly moral decision of paying €3 to ensure that the recipient receives a fixed payment of €5. Subjects play three stages, each of which consists of one instance of the 1-Dictator, the 2-Dictator, and the 3-Dictator games.[2]

In Stage 1, subjects are randomly assigned to groups and have no information about each other. I replicate the BE: subjects are more likely to make the moral decision in the 1-Dictator game than in the 2-Dictator and 3-Dictator games. More importantly, I label those subjects who make the moral decision in the 1-Dictator game as *moral dictators* and those who do not as *immoral dictators.* (The language is kept neutral in the experiment.)

Stage 2 tests whether dictators are less likely to make the moral decision if there is a high proportion of moral dictators in their group. I divide subjects into two *pools* of six

[2] As pre-registered, and to answer the question whether one individual alone or a group is more likely to produce a moral outcome, the analysis compares the 1-Dictator game to the 2-Dictator and 3-Dictator games, and not the 2-Dictator to the 3-Dictator game. The reason for using two groups of different sizes is to make sure that the results do not hold only for one specific group size.

subjects each: the *moral pool* and the *immoral pool.* The moral pool contains mostly moral dictators and the immoral pool contains mostly immoral dictators (but at least one moral one). Subjects in each pool learn that they will be assigned to groups with each other.

In line with the model, I show that the composition of the pool does not affect subjects' behavior in the 1-Dictator game. This indicates that pool composition does not alter subjects' preferences or norm perceptions, alleviating concerns that group identity in itself may alter subjects' behavior (e.g., Charness et al. 2007 and Eckel and Grossman 2005). However, as predicted, moral dictators assigned to the moral pool are *less* likely to make the moral decision than those assigned to the immoral pool in the 2-Dictator and 3-Dictator games. This implies that subjects share their responsibility to a higher degree in the moral pool. Moreover, I find that in moral pools, the moral outcome is *more* likely to be produced in the 1-Dictator than in the 2-Dictator and 3-Dictator games, a phenomenon I call the *group bystander effect* (GBE). In immoral pools, the moral outcome is *less* likely to be produced in the 1-Dictator than in the 2-Dictator and 3-Dictator games. Stage 3 replicates these results using the strategy method to elicit subjects' decisions depending on the number of moral dictators in their pool.

In order to assess the external validity of these findings, I relate the above results to the meta-analysis data of Fischer et al. (2011) on the BE, with which they investigate the probability that individuals alone or in a group help a person in need. For each of the 80 analyzed studies for which the GBE can be computed, I calculate the GBE as the probability that one individual helps minus the probability that at least one individual in a group helps. I then explore the relation between the GBE and the probability that one individual helps when alone (which proxies the proportion of moral individuals in the situation). I show that when this probability is small, the GBE is negative (i.e., people in need are more likely to be helped by groups). However, when the probability that one individual alone helps is sufficiently high, the GBE becomes positive. Coincidentally, the GBE is zero when the probability that one bystander helps is about 0.8 (as estimated by a linear regression), which is close to the experimental results in this paper. Hence, despite the stylized structure of the model and the lab experiment, I show that the model's predictions do well at explaining the existing experimental results in the field.

The main contribution of this paper is to provide a rule that can explain *when* a moral outcome is more (or less) likely to occur when there is only one individual than when there is a group. On the one hand, free-riding on others makes individuals in groups less likely to make moral decisions. On the other hand, larger groups are more likely to have at least one moral individual who is willing to make the moral decision. The former effect dominates the latter in situations in which the proportion of moral individuals is high. The latter dominates the former when the proportion of moral individuals is low. This result helps structure the mixed findings in the previous literature, in which the moral outcome is sometimes more likely to be produced by a group (e.g., Barron and Yechiam 2002; Bergstrom et al. 2019; Chekroun and Brauer 2002; Clark and Word 1974; Fromell et al. 2017; Harari et al. 1985; Shotland and Heinold 1985; Staub 1970) and others by one individual alone (e.g., Dana et al. 2007; Fischer et al. 2006; Gaertner 1975; Ross and Braband 1973; Smith et al. 1972; Shaffer et al. 1975; Van Den Bos et al. 2009).

The rule is not only useful to explain the large experimental evidence on moral decisions in groups, but may also be applied to better design organizations and institutions. One example is mentoring, which is typically a one-to-one matching between senior employees and new or less experienced employees.[3] These programs allow less experienced employees to receive help from their mentors whenever they are in need. The results in this paper indicate that, in case that the proportion of mentors who cannot help or are not willing to help is large, groups with more mentors per mentee may increase the probability that the mentee is helped. Another example is whistleblowing on bad practice in a company. Whistleblowers often risk retaliation, but their motivation to whistleblow is usually a moral one (see e.g. Near 1996). The conclusions of this paper imply that in environments where few workers are willing to bear the costs of retaliation, a group of witnesses will be more likely to report malpractice than one witness alone. However, in environments in which most workers may be willing to report malpractice, one witness will be more likely to report it than a group.

The paper proceeds as follows. Section 2 describes the related literature. Section 3 introduces a simple theoretical model used to guide my empirical investigation. Section

[3] Such programs are very common among large firms. See for example Hegstad and Wentling (2004), who discuss that 71% of Fortune 500 companies use mentoring programs.

4 outlines the experimental design used to test the hypotheses. Section 5 presents the experimental results. Section 6 relates the model and the experimental results to field data on the BE. Section 7 discusses the results and presents some conclusions.

2 Related literature

This paper relates to a growing literature in economics and psychology that studies moral decisions individually and in groups. Through a variety of experiments, this stream of research provides robust evidence that individuals are less likely to make moral decisions when they are in groups. In psychology, Latané and Nida (1981) and Fischer et al. (2011) review over two hundred experiments on the BE to find that bystanders in groups are less likely to help in 89% and 70% of the studies, respectively. Among other explanations, they argue that people feel less responsible for helping when others are present. While these reviews focus on whether people in groups are less likely to help, much of the focus has recently shifted towards the outcome. Philpot et al. (2020) for example write: "[from the perspective of the victim] the aggregated likelihood that at least someone will help [...] remains the most important question—will I receive help if needed?." The authors then use cross-country data from surveillance cameras and find that on average people in need are more likely to be helped by larger groups. The present paper shows both theoretically and empirically that whether victims are more likely to be helped by groups or individuals has a more nuanced answer: it depends on the distribution of moral preferences.

Using incentivized experiments, a large literature in economics also finds that individuals in groups behave more selfishly. For example, in situations in which a moral outcome—saving a mouse—occurs if every member of a group makes a costly moral decision, Falk et al. (2020) find that subjects are more likely to kill a mouse if they are in a group.[4] In line with these results, Panchanathan et al. (2013) and Fromell et al. (2017) show that dictators give less when other dictators can also give. Similarly, Behnk et al. (2017) use sender-received games

[4] In a similar experiment in which subjects decide whether to donate to enable a surgical operation for leprosy in India, Bartling and Özdemir (2017) do not find that subjects in groups are less likely to donate. They argue that such a diffusion of responsibility, which in their game they call *replacement logic*, may not persist when social norms are very strong.

and find that two senders acting together behave less morally towards the receiver than one sender alone. In games in which a pro-social allocation occurs as long as one dictator makes a moral decision, Dana et al. (2007) and Bergstrom et al. (2019) show that dictators are less likely to pay a cost to increase a recipient's payment when they are in a group. Similar results have been found in games where subjects in groups decide jointly whether to implement a moral decision (Bornstein and Yaniv 1998; Cox 2002; Kocher et al. 2018; Luhan et al. 2009; see also Charness and Sutter 2012 for a review comparing individual and group behavior).[5]

The present paper adds to this literature by showing that group composition is a crucial determinant of such immoral behavior in groups. More concretely, individuals who expect there to be a higher proportion of moral individuals in their group are *less* likely to make the moral decision. This somewhat counterintuitive result stems from a coordination problem in which individuals free-ride on each other hoping that another moral individual takes the responsibility.

Finally, the paper also relates to the literature on public good provision and group size. Whether and how group size affects public good provision remains an open question, with some arguing for negative effects (e.g., Baland and Platteau 1996 and Olson 1965), others for positive effects (e.g., Chamberlin 1974, Isaac and Walker 1988, Isaac et al. 1994, and Zhang and Zhu 2011), and others for non-linear and more nuanced effects (e.g., Nosenzo et al. 2015, Oliver and Marwell 1988, and Yang et al. 2013). In sum, as Nosenzo et al. (2015) and Oliver and Marwell (1988) suggest, the direction of the effect of group size on public good provision likely depends on the game played and its parameters.

This paper focuses on the threshold public good game in which the public good is provided if at least one individual contributes (Palfrey and Rosenthal 1984), also called the volunteer's dilemma (Diekmann 1985), which is one of the workhorse models for studying public provision.[6] While previous research has considered agents with heterogeneous pref-

[5] There are also a handful of papers that use theory to study the determinants of moral behavior in groups, including moral costs and group size in a committee (Huck and Konrad 2005), the distribution of decision-making power in a committee (Maaser and Stratmann 2019), and the guilt of making an immoral decision in a threshold public good game (Rothenhäusler et al. 2018).

[6] There is a large literature studying the volunteer's dilemma both theoretically and empirically. See e.g. Diekmann (1985) for a basic one-shot model with complete information, and Bliss and Nalebuff (1984) and Weesie (1993) for dynamic models with heterogeneous agents and incomplete information (see also Bergstrom 2017 for a similar model in which decisions are made sequentially). For experimental studies, see

erences (e.g., Weesie 1993 and Bliss and Nalebuff 1984), to the best of my knowledge the literature has not considered the case in which, for some individuals, the cost of producing the public good is higher than its individual benefit. By filling this gap, the present model can account for the experimental results in which sometimes a group of individuals is more likely to volunteer than one individual alone. This extension yields novel implications not only for the study of the BE and the GBE but also for the study of volunteering more generally. For example, it predicts that whenever the proportion of agents who cannot produce the public good (or whose preferences are such that they would never produce it) is sufficiently high (low), a single agent alone will be less (more) likely than a group to produce it. This is a novel finding that could have applications in some of the areas where the volunteer's dilemma has been applied, such as market entry (Sherman and Willett 1967) and voting behavior (Brennan and Lomasky 1997).

3 The model

This section presents a simple model to guide my empirical investigation. Note that the model below does not attempt to provide a substantial theoretical contribution, it is rather designed to capture the main mechanisms of interest in this paper. Online Appendices A and B extend the model to show that the results hold even in more general settings.

The first part of this section describes the setup of the model. In the second part, I analyze the symmetric Bayesian equilibrium of the game. The third part derives testable hypotheses from the equilibrium analysis. Finally, the fourth part discusses the model's limitations and potential extensions.

3.1 Setup

Every agent in a group of $n \geq 1$ agents decides simultaneously whether to make a moral decision (play M) or not (play $\neg M$). Hence, a (mixed) strategy for an agent specifies the probability that the agent plays M. The moral outcome occurs if and only if at least one

e.g. Diekmann (1993), Franzen (1995), Goeree et al. (2017), Hillenbrand and Winter (2018), and Hillenbrand et al. (2020) for one-shot experiments and Otsubo and Rapoport (2008) and Babcock et al. (2017) for dynamic experiments.

agent plays M. Agents pay a cost c for making the moral decision, and receive a payoff b_i if the moral outcome occurs. These parameters are fixed and do not depend on n. In what follows, I assume that there are two types of agents: moral agents ($i = m$), for whom $b_m > c$, and immoral agents ($i = im$), for whom $b_{im} < c$. Agents know their own type but they do not know the type of the other $n - 1$ agents in their group. However, they know the proportion γ of moral agents in the population.[7]

Denote by $p(n, \gamma)$ the expected probability that a randomly selected agent plays M when in a group of $n \geq 1$ agents and the proportion of moral agents in the population is γ.

Definition 1. *The Bystander Effect (BE) is the probability that the agent makes the moral decision when in a group of one as compared to being in a group of* $n > 1$ *agents:*

$$BE \equiv p(1, \gamma) - p(n, \gamma)$$

Furthermore, let $P(n, \gamma)$ be the probability that the moral outcome occurs in a group of $n \geq 1$ agents when the proportion of moral agents in the population is γ.[8]

Definition 2. *The Group Bystander Effect (GBE) is the probability that the moral outcome occurs in a group of one agent as compared to in a group of* $n > 1$ *agents:*

$$GBE \equiv P(1, \gamma) - P(n, \gamma)$$

Hence, while the BE captures how each agent's decision is affected by being in a group rather than alone, the GBE captures how the *aggregate outcome* changes when there is a group rather than one agent alone. The analysis below studies the BE, the GBE, and how they are affected by γ.

[7] Section 3.4 discusses that assuming that b_i decreases in n, as some previous experimental papers suggest, does not meaningfully change the results. Online Appendix A adds warm-glow and conformity preferences and Online Appendix B extends the model by assuming that there is a continuum of types that differ in their valuation of the moral outcome b_i. The main results of the paper hold in these cases.

[8] Note that $P(n, \gamma) = 1 - (1 - p(n, \gamma))^n$.

3.2 Equilibrium

For immoral agents, the cost of making the moral decision is, by definition, always higher than the benefit that they get if the moral outcome occurs. Therefore, regardless of n, immoral agents always play $\neg M$.

Next consider moral agents and suppose $n = 1$. In this case, the model boils down to an individual decision whether to make the moral decision or not. Since $b_m > c$ by definition, every moral agent plays M. Therefore, since a proportion γ of the agents is moral, on average agents play M with probability γ and thus $p(1, \gamma) = \gamma$. When $n = 1$, the moral outcome can only be produced by one agent, hence $P(1, \gamma) = \gamma$.

When $n > 1$, there may exist multiple pure strategy equilibria. As usual in the volunteer's dilemma literature (Diekmann 1985), I will focus the analysis on the unique symmetric Bayesian equilibrium. Note that this is the most reasonable equilibrium in real-world situations where agents are strangers to one another or are not able to coordinate on pure strategy equilibria. In this equilibrium, moral agents play M with positive probability and immoral agents always play $\neg M$.

Proposition 1. *Let $\sigma^* = \frac{1}{\gamma}\left(1 - \left(\frac{c}{b_m}\right)^{\frac{1}{n-1}}\right)$. In the unique symmetric Bayesian equilibrium, moral agents play M with probability*

$$\sigma_m(n, \gamma) = \begin{cases} \sigma^* & \text{if } \gamma > 1 - \left(\frac{c}{b_m}\right)^{\frac{1}{n-1}} \\ 1 & \text{if } \gamma \leq 1 - \left(\frac{c}{b_m}\right)^{\frac{1}{n-1}} \end{cases}$$

Proof. In equilibrium, moral agents only mix between M and $\neg M$ if they are indifferent between both strategies, i.e. $EU(M) = EU(\neg M)$ or equivalently $b_m - c = (1 - (1 - \gamma\sigma_m(n, \gamma))^{n-1})b_m$. Solving this equation for $\sigma_m(n, \gamma)$ yields the desired result

$$\sigma_m(n, \gamma) = \sigma^*$$

Since $\sigma_m(n, \gamma) \in [0, 1]$, then whenever $\gamma < 1 - \left(\frac{c}{b_m}\right)^{\frac{1}{n-1}}$ the above equality does not hold. Instead, $EU(M) > EU(\neg M)$, implying that moral agents strictly prefer playing M over $\neg M$ when the proportion of moral agents is sufficiently low. □

The analysis below uses this proposition to derive results in the form of hypotheses regarding the BE and the GBE. I then contrast these predictions with the experimental data in Section 5.

3.3 Predictions

First, note that moral agents always make the moral decision when they are alone, but they may or may not always make it when they are in a larger group. Thus, the probability that a randomly selected agent makes the moral decision when alone is $p(1, \gamma) = \gamma$ (the probability of being moral), and the probability of making the moral decision when in a group of $n > 1$ agents is $p(n, \gamma) = \gamma\sigma_m(n, \gamma)$. Hence, $BE = \gamma - \gamma\sigma_m(n, \gamma) = \gamma(1 - \sigma_m(n, \gamma)) \geq 0$.

Hypothesis 1. *Agents are (weakly) more likely to make the moral decision when they are alone than when they are in a group, i.e., $BE \geq 0$ for all n and γ.*

How do agents' behavior, the BE, and the GBE react to changes in γ? Consider first the probability that a moral agent makes the moral decision. When $n = 1$, moral agents always play M regardless of γ. When $n > 1$, moral agents play M with probability $\sigma_m(n, \gamma)$. Note that σ^* is decreasing in γ. Thus, when $\sigma_m(n, \gamma) = \sigma^*$ agents are less likely to play M as γ increases. When $\sigma_m(n, \gamma) = 1$, agents are weakly less likely to play M as γ increases.[9]

Hypothesis 2. *Moral agents in groups are (weakly) less likely to make the moral decision when the expected proportion of moral agents in their group is higher, i.e., $\frac{\partial \sigma_m(n,\gamma)}{\partial \gamma} \leq 0$ for all n and γ.*

Furthermore, since $BE = \gamma - \gamma\sigma_m(n, \gamma) = \gamma(1 - \sigma_m(n, \gamma))$, it follows that the BE is increasing in γ.

Note that the probability that the moral outcome occurs when $n = 1$ is $P(1, \gamma) = p(1, \gamma) = \gamma$, and the probability that it occurs in a group of $n > 1$ agents is $P(n, \gamma) = (1 - (1 - \gamma\sigma_m(n, \gamma))^n)$. Thus, $GBE = \gamma - (1 - (1 - \gamma\sigma_m(n, \gamma))^n)$. Hence, whether GBE is positive or negative depends on the particular values of the parameters b_m, c, n and γ.

[9] If γ increases enough such that at some point $\sigma_m(n, \gamma) < 1$, then agents are less likely to play M with higher γ. If γ does not increase enough such that $\sigma_m(n, \gamma) < 1$, then agents are equally likely to play M with higher γ.

Proposition 2. *For every $n > 1$, there exists a $\gamma^* \in (0, 1)$ such that the moral outcome is equally likely to occur when there is one agent alone as to when there is a group of n agents, i.e. $GBE = 0$. If $\gamma > \gamma^*$, the moral outcome is more likely to occur when there is one agent alone, i.e. $GBE > 0$. If $\gamma < \gamma^*$, the moral outcome is more likely to occur when there is a group of n agents, i.e., $GBE < 0$.*

Proof. Note that if γ is low enough, such that $\sigma_m(n, \gamma) = 1$, then $GBE = \gamma - 1 + (1 - \gamma)^n < 0$, which is negative since $(1 - \gamma)^n < 1 - \gamma$. Furthermore, if $\gamma = 1$, then note by Proposition 1 that $\sigma_m(n, \gamma) < 1$. Hence, $(1 - \sigma_m(n, \gamma))^n > 0$ which implies that $GBE = 1 - (1 - (1 - \sigma_m(n, \gamma))^n) > 0$. Now define γ^* such that $GBE = 0$. Since $GBE < 0$ when γ is such that $\sigma_m(n, \gamma) = 1$, it must be that $\sigma_m(n, \gamma) < 1$ for $GBE = 0$. Since $GBE = \gamma^* - (1 - (1 - \frac{b_m - c}{b_m})^{\frac{n}{n-1}}) = 0$, there exists a unique $\gamma^* = 1 - (1 - \frac{b_m - c}{b_m})^{\frac{n}{n-1}}$ for which $GBE = 0$. □

The above proposition yields the following testable hypothesis:

Hypothesis 3. *If γ is sufficiently high, the moral outcome is more likely to occur when there is one agent alone, i.e., $GBE > 0$ for all n. If γ is sufficiently low, the moral outcome is more likely to occur when there is a group, i.e., $GBE < 0$ for all n.*

3.4 Limitations and extensions

The stylized model presented above necessarily abstracts from some potentially relevant features. In particular, three modeling choices may need some further discussion.

First, the model predicts that agents in groups are less likely to make the moral decision because they strategically interact with each other. While strategic interaction has often been used as an explanation of diffusion of responsibility in groups (e.g., Diekmann 1985, Weesie 1993, and Campos-Mercade 2018b), such a consequentialistic approach is not enough to explain some of the evidence for diffusion of responsibility. In Dana et al. (2007), for example, dictators act as if they had fewer moral concerns for the recipient when they are in a group than when they are alone. In reality, when the action of one agent is needed for a moral outcome to be produced, it is likely that both strategic interaction and diminishing moral concerns in groups play a role in the observed diffusion of responsibility. Introducing such diminishing moral concerns in the form of a reduced $b'_m < b_m$ when $n > 1$ would trivially not change any of the results as long as $b'_m > c$. However, if $b'_m < c$, then moral agents in groups would never make the moral decision. In that case, $p(n, \gamma) = 0$ and $P(n, \gamma) = 0$ when $n > 1$.[10]

Second, there is evidence that individuals are often driven by non-consequentialistic preferences such as warm-glow and conformity preferences, which are not captured in the model. Indeed, there may exist situations in which people make moral decisions because it feels good to make them, or because they have a desire to conform to the norm. One might worry that the model could lose its predictive power in such situations. To deal with this concern, Online Appendix A extends the model by allowing agents to have warm-glow preferences (they receive additional utility just by making the moral decision, as in Andreoni 1990) and to also care about conforming to the norm (they receive positive utility if they make the decisions that they expect others to make, similar to Bernheim 1994). As long as agents care to some extent about whether the moral outcome occurs, the model yields very similar results to those in the main text.[11]

Third, the model assumes that the proportion of moral agents can change independently of the other parameters, without acknowledging that this proportion could be endogenous to the situation. For example, in situations where the cost of making the moral decision is low, it is likely that the proportion of agents who would be willing to make the moral decision (the moral agents) is higher than in situations where this cost is high. To deal with this concern, Online Appendix B extends the model by endogenizing the proportion of moral agents. It assumes that there is a continuum of types that differ in how much they value the moral outcome. Under the assumption that the cumulative distribution function of the moral outcome value is S-shaped (such as the normal distribution), I show that the GBE

[10] While theoretically possible, note that this extreme outcome is not common in previous empirical evidence. In Fischer et al. (2011), for example, only 2 of the 91 studies for which $p(n, \gamma)$ can be computed find $p(n, \gamma) = 0$.

[11] In the extreme cases in which agents make the moral decision almost only due to warm-glow or conformity preferences, such results are lost. If warm-glow preferences were sufficiently strong, then moral agents would always make the moral decision regardless of the group size and the rest of the parameters. If conformity preferences were sufficiently strong, then agents would always do what they expect the rest to do, regardless of the other parameters. While theoretically possible, both of these cases would be hard to reconcile with the strong evidence for diffusion of responsibility and the BE.

is positive for low costs of making the moral decision (i.e., when most agents are willing to make it) and negative for high costs of making the moral decision. These results are analogous to those obtained in the discrete case with only two types of agents.[12]

4 Experimental design

The experimental data were collected at the WISE lab at the University of Hamburg.[13] The experiment was programmed in z-Tree (Fischbacher 2007) and subjects were recruited using the hroot software (Bock et al. 2014). A total of 378 subjects participated throughout 21 sessions of 18 subjects each. Subjects participated in the experiment anonymously, using different computers, and without interacting with one another. Each session lasted approximately 60 minutes and subjects earned on average €12 (consisting of a €5 show-up fee and an average of €7 in experimental earnings).

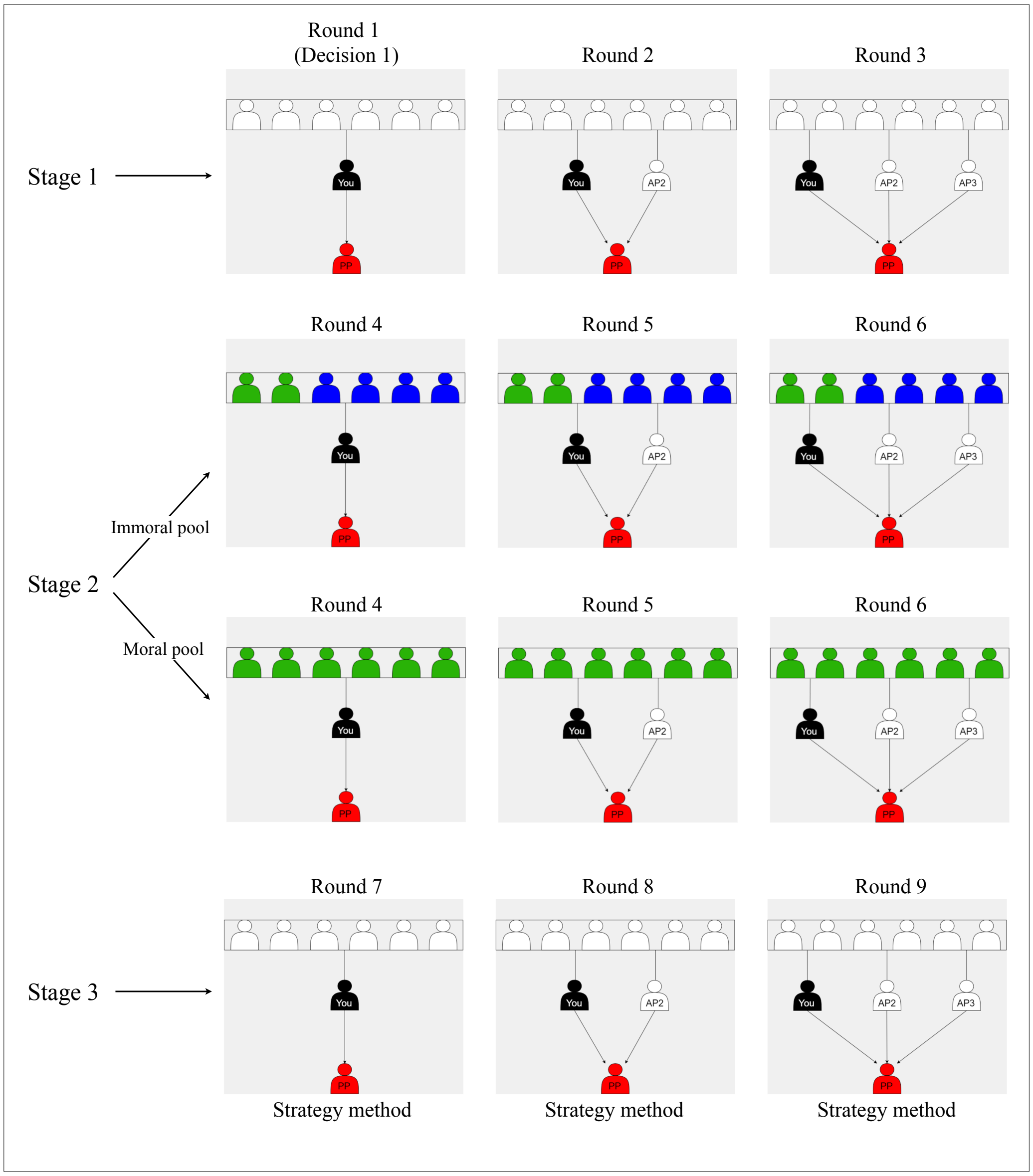
Subjects were told that they had been randomly given the role of a *dictator* (referred to as an *Active Participant*) or *recipient* (*Passive Participant*). However, during the experiment they would all make decisions as dictators and only at the end of the experiment would their role be revealed to them.

Each subject played nine rounds. They did not learn the outcome of any of the rounds until the end of the experiment, when one round was randomly chosen for payment. In each round, subjects played in groups of one recipient and either one, two, or three dictators (the 1-Dictator, 2-Dictator, and 3-Dictator games, respectively). Dictators and recipients started each round with an endowment of €10 and €0, respectively. Dictators then decided simultaneously and independently whether to *pay* or *not pay.* Paying carried a cost of €3. If *at least* one dictator in the group paid, then the recipient got a payment of €5. If none of the dictators paid, the recipient got a payment of €0. The analysis assumes that to *pay* corresponds to making the moral decision M in the model above, and to *not pay* corresponds

[12] When combining diminishing moral concerns in groups, warm-glow preferences, conformity preferences, and a continuum of types, the model's predictions are likely to not always hold in general. Importantly, however, the mechanism that drives the model's results should still hold.

[13] The experiment did not get an IRB approval because my university at the time, Lund University, did not require evaluation (nor evaluated) experiments that followed common practice procedures in experimental economics.

Figure 1. Subjects' interface during the experiment



Note: Screens that subjects saw when making their decisions in the nine rounds. The pictures display a recipient (PP) who is assigned to either one dictator or a group of two or three dictators (AP) drawn from a pool of six dictators. In Stage 1 and Stage 3, dictators did not have any information about the dictators in their pool. In Stage 2, dictators learned the number of dictators in their pool who had paid in Decision 1. Subjects who paid were colored in green, and subjects who did not pay were colored in blue. In this picture, there are two examples: an immoral pool (in which two dictators had paid in Decision 1 and four had not) and a moral pool (in which all dictators had paid in Decision 1). In Stage 1 and Stage 2, the subjects' choice in each round consisted of either paying or not paying. In Stage 3, subjects used the strategy method to choose whether to pay depending on the number of dictators in their pool who had paid in Decision 1.

to $\neg M$.

As depicted in Figure 1, the nine rounds were divided into three stages of three rounds each. Subjects received stage-specific instructions and answered a set of control questions right before playing in each stage. Subjects were aware that there would be three stages, but while playing in each stage they were unaware about the game that they would play in the future stages. In each stage, subjects played the 1-Dictator, 2-Dictator, and 3-Dictator games once and in random order.

In Stage 1, subjects were told that they had been placed in a pool of six dictators and that they would play the three games with other dictators from that pool. Stage 2 followed the same structure as Stage 1, with one key difference: subjects were placed in a new pool, and they learned how many of the six dictators in this pool (including themselves) had paid in the 1-Dictator game in Stage 1 (a decision that was labeled *Decision 1*). For exposition purposes, in what follows I refer to those subjects who paid in Decision 1 as *moral dictators* and those who did not pay as *immoral dictators.* (In the experiment, they were referred to as *participants who chose to [not] pay in Decision 1.*) In Stage 3, subjects were told that they had been reshuffled into a new pool, of which they did not know the number of moral and immoral dictators. However, instead of deciding whether to pay or not pay as in Stage 1 and Stage 2, subjects used the strategy method to decide whether to pay conditional on the number of moral dictators in their pool. They therefore answered whether to pay in a list of six decisions, from "If no other participants in my pool paid in Decision 1, I choose to" to "If five other participants in my pool paid in Decision 1, I choose to."[14,15]

The main exogenous treatment took place in Stage 2. Recall that each session consisted of 18 subjects, 12 dictators, and 6 recipients. For the 12 dictators, at the end of Stage 1 the computer counted how many of them paid in Decision 1. It then created two pools, which I refer to as the *moral pool* and the *immoral pool.* The moral pool consisted mostly of moral dictators, while the immoral pool consisted mostly of immoral dictators. The number

[14] If one of the rounds in Stage 3 was selected for payment, the computer counted the number of other moral dictators in the pool and the decision that was implemented for payment was the one that subjects made for that number of moral dictators.

[15] Additionally, Stage 1 and Stage 2 elicited subjects' beliefs about the decisions of the rest of the dictators in their pool. Subjects were told that if they correctly guessed the number of the other five dictators who paid in the round chosen for payment (where one, two, or three dictators would pay), they would, in addition, be paid €1 at the end of the experiment.

of moral and immoral dictators in each pool depended on the number of dictators in the session who paid in Decision 1. For example, if 7 of the 12 dictators in the session paid, then the moral pool would consist of 5 moral dictators and 1 immoral dictator, and the immoral pool would consist of 2 moral dictators and 4 immoral dictators.[16] In most sessions, about 6 to 9 dictators paid in Decision 1. Subjects were only told about the composition of their pool and were unaware about how such pool had been created.

For the 6 recipients, Stage 2 always showed them that they were in an immoral pool where only 1 dictator paid in Decision 1. This feature of the design increases the sample size of moral dictators in immoral pools and the statistical power to test Hypothesis 2. The experiment avoided deceiving these 6 recipients since, although they did *not* know that they were recipients and hence had incentives to answer truthfully, the instructions stated that recipients would not see real scenarios.

5 Results

This section uses the experimental data to test each of the three hypotheses proposed in Section 3. The analysis was pre-registered in the AEA RCT Registry (Campos-Mercade 2018a). Online Appendix D contains a copy of the pre-registration.

Figure 2 shows the proportion of subjects who pay for each of the three rounds of Stage 1 and their respective 95% confidence intervals. Hypothesis 1 states that dictators in a group are less likely to pay than when they are the only dictator. A McNemar test provides strong support for this hypothesis ($p < 0.001$ both when comparing the 1-Dictator to the 2-Dictator game and when comparing the 1-Dictator to the 3-Dictator game, $N = 378$ in each arm). Further restricting the analysis to subjects' first round decisions (recall that the order of the rounds is randomized within each subject), maintains the significance of the results (using a chi-square test, $p = 0.017$ when comparing the 1-Dictator to the 2-Dictator game and $p = 0.032$ when comparing the 1-Dictator to the 3-Dictator game, $N = 126$ in each arm).

[16] See Online Appendix C for the full table of how the moral and the immoral pools were created depending on the number of dictators who paid in Decision 1. Appendix C also shows how many pools of each composition were created during the experiment.

Stage 1 thus does replicate the BE.[17]

Result 1. *Dictators are more likely to make the moral decision in the 1-Dictator game than in the 2-Dictator and 3-Dictator games.*

Hypothesis 2 states that the probability that moral dictators pay when they are in a group with other dictators decreases in the proportion of moral dictators. Recall that the model defines those agents who would make the moral decision if they were the only decision-makers as “moral agents.” In the experiment, I therefore define those subjects who pay in Decision 1 as “moral dictators” and those who do not as “immoral dictators.” Since moral dictators are randomly assigned to one of the pools, I test whether moral dictators assigned to the moral pool are less likely to pay than those assigned to the immoral pool. Importantly, the model predicts this difference in the 2-Dictator and 3-Dictator games, but not in the 1-Dictator game.

The data provides support for these hypotheses. There are in total 107 moral dictators assigned to a moral pool and 124 moral dictators assigned to an immoral pool. On average, moral pools have 5.09 moral dictators and immoral pools have 1.45 moral dictators. Figure 2 represents moral dictators’ paying frequency by their pool in Stage 2, along with their 95% confidence intervals. In the 1-Dictator game, moral dictators pay about 85% of the time regardless of the pool they are assigned to. In the 2-Dictator game, 69% of those assigned to the immoral pool pay and 52% of those assigned to the moral pool pay. In the 3-Dictator game, 60% of those assigned to the immoral pool pay and only 40% of those assigned to the moral pool pay. Table 1 tests these differences using a regression analysis. It shows that the results hold when controlling for the order in which subjects play the rounds within each stage and for subjects’ decisions in Stage 1. More concretely, subjects do not behave differently if they are assigned to the moral pool in the 1-Dictator game. However, in the 2-Dictator and 3-Dictator games, they are respectively 12.4 and 16.1 percentage points less

[17] While this analysis corresponds only to Stage 1 decisions, the same pattern emerges for Stage 2 and Stage 3. Regardless of the stage, dictators are more likely to pay in the 1-Dictator game than in the 2-Dictator and 3-Dictator games (in Stage 3, I compute this proportion by taking the average paying frequency over all the conditional decisions). More concretely, when pooling all decisions across all rounds, dictators pay 56%, 43%, and 36% of the times in the 1-Dictator, 2-Dictator, and 3-Dictator games, respectively (all comparisons are statistically significant when using a t-test, $p < 0.001$).

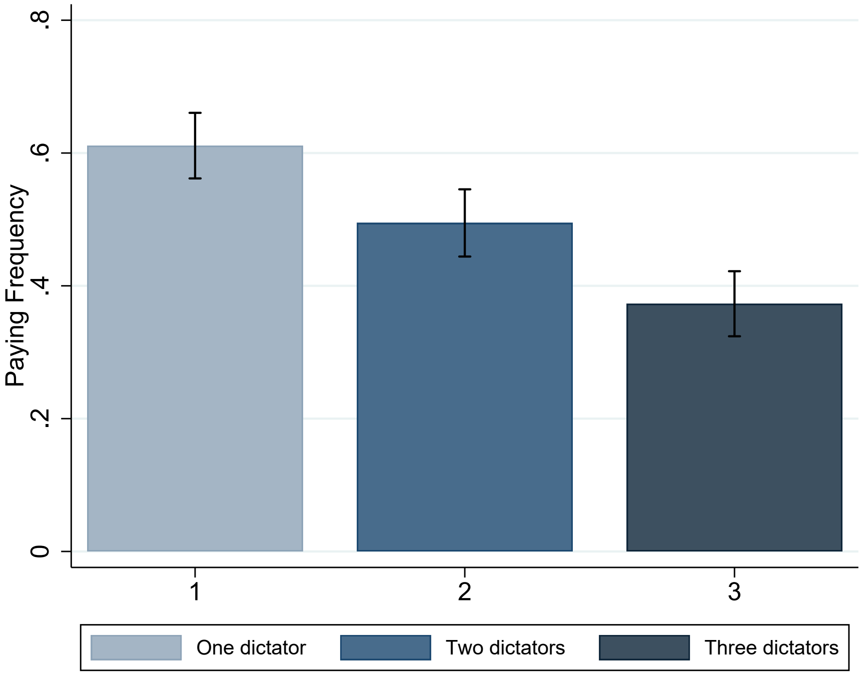


Figure 2. Paying frequency in Stage 1 by the number of dictators

likely to pay if assigned to the moral pool (the differences are significant at the $p = 0.036$ and $p = 0.006$ levels).[18]

Result 2. *In the 1-Dictator game, moral dictators in a moral pool are equally likely to make the moral decision as are those in an immoral pool. In the 2-Dictator and 3-Dictator games, moral dictators are less likely to make the moral decision when in a moral pool than when in an immoral pool.*

Result 2.1. *The BE is higher in moral pools than in immoral pools.*

Recall that Section 3 defines the BE as the probability that one agent makes the moral decision when alone minus the probability of making the moral decision when in a group. One theoretical implication of the previous statement is that the BE increases with the percentage of moral agents. Indeed, in the 2-Dictator game, I find that the BE in the immoral pool is $BE = 0.10$ and in the moral pool $BE = 0.29$. Computing the BE for each subject, which takes the value of either 1 (if the subject pays in the 1-Dictator but not in

[18] As implied by the model, in Stage 2 immoral dictators almost do not pay in any of the conditions: their paying rate across all rounds is 9.3%, which does not significantly change with the composition of the pool.

Table 1. Treatment effect on whether to pay with 2 and 3 dictators in Stage 2 (moral subjects only)

	1-D (1)	1-D (2)	2-D (3)	2-D (4)	3-D (5)	3-D (6)
Moral pool	0.004	0.024	-0.170	-0.124	-0.203	-0.161
	(0.047)	(0.043)	(0.064)***	(0.059)**	(0.065)***	(0.058)***
2 dictators		0.188		0.294		0.201
(Stage 1)		(0.060)***		(0.081)***		(0.075)***
3 dictators		0.070		0.315		0.461
(Stage 1)		(0.045)		(0.072)***		(0.070)***
Constant	0.847	0.536	0.694	0.374	0.605	0.188
	(0.032)***	(0.087)***	(0.042)***	(0.099)***	(0.044)***	(0.097)*
Order FE	NO	YES	NO	YES	NO	YES
Session FE	NO	YES	NO	YES	NO	YES
R^2	0.00	0.26	0.03	0.32	0.04	0.37
N	231	231	231	231	231	231

Note: This table reports the results of an OLS regression to test the effect of being in a moral pool rather than in an immoral pool. The analysis is restricted to moral subjects. The outcome variable takes value 1 if the subject pays in that round and 0 otherwise. 1-D, 2-D, and 3-D are dummies that take value 1 if the subject pays in the 1-Dictator, 2-Dictator, and 3-Dictator game, respectively. Moral pool is a dummy that takes value 1 if the subject is assigned to the moral pool. Two dictators (Stage 1) and Three dictators (Stage 1) are dummies that take value 1 if the subject paid in the 2-Dictator and 3-Dictator games in Stage 1, respectively. Order FE and Session FE include dummies controlling for the six possible orders in which subjects played the different rounds and the twenty-one sessions. Robust standard errors are reported in parentheses. OLS is used for an easier interpretation of the coefficients, but logit and probit models yield the same results in terms of significance. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

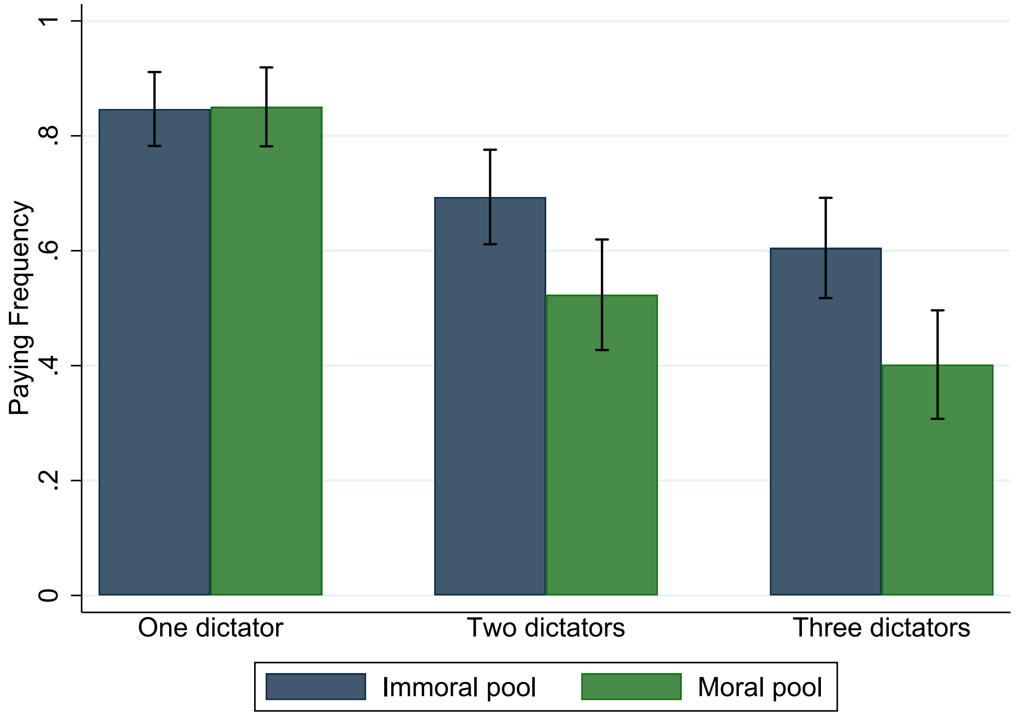


Figure 3. Paying frequency of moral dictators in Stage 2 by the composition of their pool

the 2-Dictator game), 0 (if the subject makes the same decision in the 1-Dictator as in the 2-Dictator game), or -1 (if the subject does not pay in the 1-Dictator game but does in the 2-Dictator game), the difference is significant at the $p < 0.001$ level using a standard t-test ($N = 126$ for the moral pool and $N = 252$ for the immoral one). In the 3-Dictator game, using the same analysis, the BE is $BE = 0.13$ in the immoral pool and $BE = 0.38$ in the moral pool ($p < 0.001$).

Hypothesis 3 changes the perspective to the aggregate outcome. It states that the GBE—defined as the probability that the moral outcome occurs when there is one dictator minus the probability that it occurs when there are several dictators—is positive when the proportion of moral dictators is sufficiently high, but negative for a lesser proportion of moral dictators. To test this hypothesis, I use data from Stage 3, where subjects answer whether to pay using the strategy method conditional on how many moral dictators there are in their pool. I first test whether the recipient is more likely to obtain the payment in the 1-Dictator game than in the 2-Dictator and 3-Dictator games when the pool consists of six moral dictators.

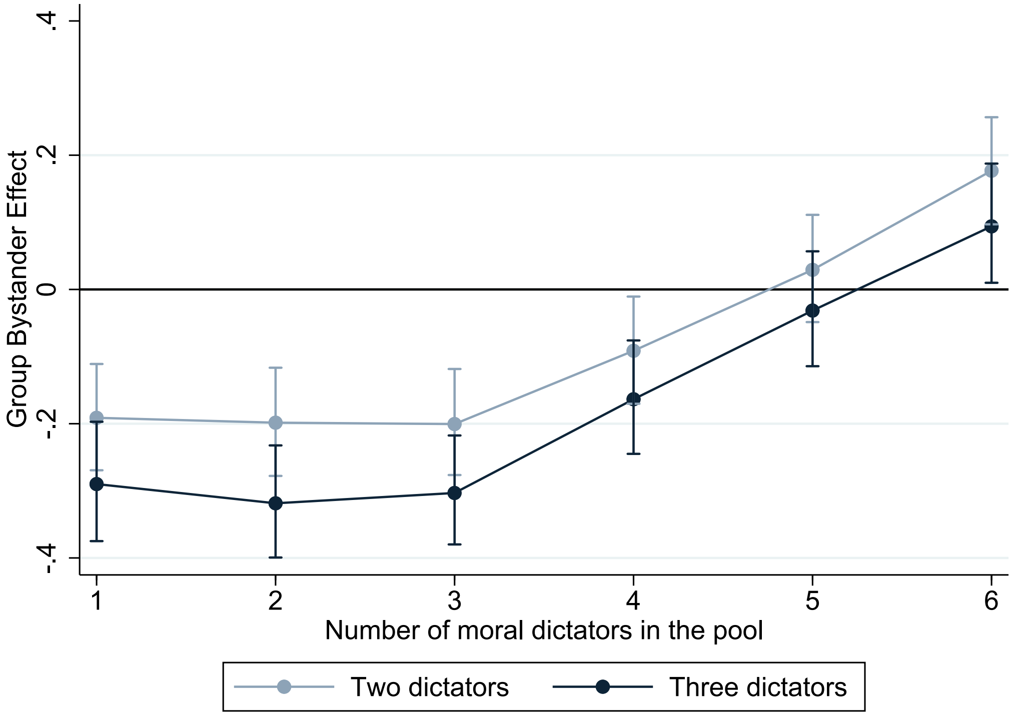


Figure 4. GBE based on dictators' decisions in Stage 3 by the number of moral dictators in their pool

I then test whether the recipient is more likely to obtain the payment in the 2-Dictator and 3-Dictator games than in the 1-Dictator game when the pool has only one moral dictator.

Figure 4 shows the computed GBE both for two and three dictators as compared to one dictator using the decisions in Stage 3 (the error bars show the 95% confidence intervals). To compute the GBE for each comparison ($n = 2$ and $n = 3$ as compared to $n = 1$) and each pool composition $\gamma \in \{\frac{1}{6}, ..., \frac{6}{6}\}$, I use the frequencies $p(1, \gamma)$ and $p(n, \gamma)$ with which subjects choose to pay in each case. For example, when the number of moral dictators in the pool is five (i.e., $\gamma = \frac{5}{6}$), moral dictators pay with probability 0.69 in the 1-Dictator game and 0.39 in the 2-Dictator game. In this case, immoral dictators pay with probability 0.12 in the 1-Dictator game and 0.09 in the 2-Dictator game. Hence, to compute the GBE for two dictators, note that $p(1, \gamma) = \frac{5}{6}0.69 + \frac{1}{6}0.12 = 0.595$ and $p(2, \gamma) = \frac{5}{6}0.39 + \frac{1}{6}0.09 = 0.34$ (the probability that a given dictator is moral/immoral times the probability with which this kind of dictator pays). Hence, when the pool has five moral dictators, the GBE for two dictators as compared to one is $GBE = p(1, \gamma) - (1 - (1 - p(n, \gamma))^n) = 0.595 - (1 - (1 - 0.34)^2) = 0.03$, as displayed in Figure 4.

Note that the GBE does not obey any of the standard distributions. Hence, to compute the confidence intervals and perform the tests, I bootstrap the dictators' decisions in Stage 3.[19] When there is only one moral dictator in the pool (i.e., $\gamma = \frac{1}{6}$), then $GBE = -0.19$ for $n = 2$ and $GBE = -0.29$ for $n = 3$ (which are statistically different from zero, $p < 0.001$ for both tests). This means that recipients are more likely to be paid in the 2-Dictator and 3-Dictator games than in the 1-Dictator game. When there are six moral dictators in the pool (i.e., $\gamma = \frac{6}{6}$), then $GBE = 0.18$ for $n = 2$ and $GBE = 0.09$ for $n = 3$ ($p < 0.001$ and $p = 0.038$, respectively), which means that recipients are more likely to be paid in the 1-Dictator game than in the 2-Dictator and 3-Dictator games. In Stage 2, with much fewer observations and therefore less power to perform the statistical tests, the pattern is the same. For $\gamma = \frac{1}{6}$, $GBE = -0.07$ for $n = 2$ and $GBE = -0.21$ for $n = 3$. For $\gamma = \frac{6}{6}$, $GBE = 0.14$ for $n = 2$ and $GBE = 0.04$ for $n = 3$.

Note also that, using a simple linear interpolation between the data points, there is a unique proportion γ^* of moral dictators in the pool for which the GBE is zero, both in the case of two dictators and that of three dictators. This γ^* is around the pool consisting of five moral dictators, or $\gamma^* \approx \frac{5}{6} = 0.83$.

Result 3. *When the proportion of moral dictators is high, the moral outcome is more likely to occur in the 1-Dictator game than in the 2-Dictator and 3-Dictator games. When the proportion of moral dictators is low, the moral outcome is more likely to occur in the 2-Dictator and 3-Dictator games than in the 1-Dictator game.*

In short, the experiment yields three key results supporting the model's hypotheses. First, subjects are less likely to make a costly moral decision when they are in a group with

[19] Technically, to create the confidence interval for the GBE, I bootstrap subjects' decisions and compute the GBE 10,000 times for each comparison ($n = 2$ and $n = 3$ as compared to $n = 1$) and each pool composition $\gamma \in \{\frac{1}{6}, ..., \frac{6}{6}\}$. I then sort the 10,000 results from lowest to highest and pick the numbers in positions 250 and 9750 as the lower and upper bounds of the confidence interval. To compute the p-values of the GBE for $\gamma = \frac{1}{6}$ and GBE for $\gamma = \frac{6}{6}$, I use a similar method. I bootstrap subjects' decisions in the 1-Dictator game, compute the average bootstrapped frequency of paying $p'(1, \gamma)$, and compute the $p'(n, \gamma)$ for which $GBE' = p'(1, \gamma) - p'(n, \gamma) = 0$. To allow for noise, I assume that subjects pay according to a binomial distribution with probability of success equal to $p'(n, \gamma)$ and compute the average simulated $p''(n, \gamma)$ such that $GBE'' = p'(1, \gamma) - p''(n, \gamma) = 0$. I repeat the process 10,000 times. To infer the p-value of the test $GBE = 0$, I sort the 10,000 results for $p''(n, \gamma)$ from lowest to highest and check where the experimental $p(n, \gamma)$ lies.

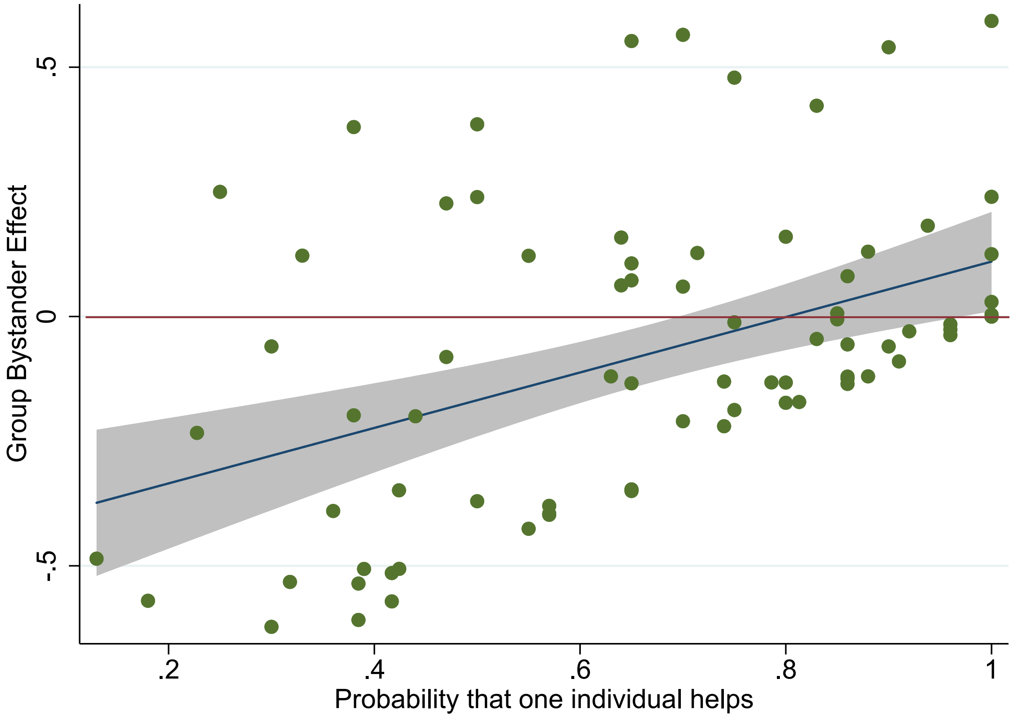


Figure 5. GBE of each study in the meta-analysis of Fischer et al. (2011), by the probability that one individual helps

other subjects. Second, moral subjects in groups are less likely to make the moral decision when the expected proportion of moral dictators is higher. Third, the moral outcome is more likely to be produced by one dictator in situations with a high proportion of moral dictators, and by two or three dictators in situations with a low proportion of moral dictators.

6 Explaining the experimental evidence about the bystander effect

This paper yields the conclusion that moral outcomes are more likely to be produced by one individual alone in situations in which the proportion of moral individuals is high, while they are more likely to be produced by a group in situations in which the proportion of moral individuals is low. In this section, I confront this prediction with data from Fischer et al. (2011), the latest meta-analysis of the BE, who study how the probability that individuals help a person in need changes depending on whether they are alone or in a group. In this

Table 2. Regressing the GBE on the probability that one individual helps using the data in Fischer et al. (2011)

	GBE	GBE	GBE Positive	GBE Positive
	(1)	(2)	(3)	(4)
γ	0.556	0.616	0.500	0.558
	(0.123)***	(0.119)***	(0.208)**	(0.204)***
# Individuals		-0.079		-0.076
		(0.023)***		(0.055)
Constant	-0.446	-0.360	0.051	0.133
	(0.098)***	(0.101)***	(0.144)	(0.166)
R^2	0.20	0.26	0.06	0.08
N	80	80	80	80

Note: This table reports the results of an OLS regression that explains the GBE of each study in Fischer et al. (2011) with the probability that one individual helps in that study. For each observation, the GBE is computed as the probability that help is provided when there is one individual minus the same probability when there is a group of individuals. The dependent variable in columns (1) and (2) is the GBE, and the dependent variable in columns (3) and (4) is a binary variable that takes the value 1 if the GBE is positive and 0 if it is negative. γ is the probability that one individual helps when alone. # Individuals is the number of individuals present in the treatment with a group of individuals. Robust standard errors reported in parentheses. $***, **, *$ indicate statistical significance at the 0.01, 0.05, and 0.10 levels using two-tailed tests.

case, I consider the decision to help as the moral decision, and that the person in need is helped as the moral outcome. I compute the GBE for the 80 studies for which it can be computed and study how the GBE depends on the probability that one individual helps.[20]

Figure 5 studies the relation between the GBE and the probability that one individual helps. Each dot in the graph represents the outcome of one study, and the study's probability that one individual helps is plotted against the study's GBE. The blue line is a linear fit to the data, and the gray area is its 95% confidence interval. Note that, in line with the model, the GBE is negative when the probability that one individual helps is low, and is positive

[20] While this analysis is important to understand whether this paper's predictions are in line with field data, one should be cautious with its interpretation. In this analysis, the probability that one individual helps (the exogenous variable) is used to estimate the GBE (the endogenous variable). Hence, since noise in the probability that one individual helps is fully absorbed by the GBE variable, one should expect the relation between both variables to be positive, even in the absence of an effect. (Note that the experiment above side-steps this problem by using different decisions to calculate the proportion of moral agents and the GBE.)

when the probability that one individual helps is high. A linear fit of the data shows that the turning point is around 80% (which, coincidentally, is similar to the $\gamma^* \approx \frac{5}{6} = 83\%$ found in the lab experiment).

To test this relation, Table 2 explains both the GBE and whether the GBE is positive with the probability that one individual helps when alone for each of the studies in Fischer et al. (2011). Columns (1) and (2) show that there is a positive correlation between the probability that one individual helps when alone (γ) and the GBE ($p < 0.001$ in column 2). Furthermore, columns (3) and (4) show that there is also a positive relation between γ and whether the GBE is positive ($p = 0.008$ in column 4).

7 Conclusion

In many situations, a moral outcome occurs if one agent makes a costly decision. In this paper, I propose and test a model to explain whether it is more likely that the moral outcome is produced by one agent alone or by a group of agents. The model assumes that agents are either moral (i.e., they would be willing to make the decision if they knew that no one else would) or immoral. I show that 1) in situations with a high proportion of moral agents, the moral outcome is more likely to be produced by one agent, whereas 2) in situations with a low proportion of moral agents, the moral outcome is more likely to be produced by a group.

The model in this paper may be too simple to capture some important nuances of the real world when agents decide whether to make a moral decision. A first limitation is that it assumes that agents can be classified into moral types who merely care about whether a moral outcome occurs. In this setting, the fact that agents in groups are less likely to make moral decisions can be solely explained by beliefs: agents in groups are less willing to make the moral decision because they believe that with some probability some other agent will. Such a consequentialistic approach is, however, not sufficient to explain some of the existing evidence on diffusion of responsibility (e.g. Dana et al. 2007, Cryder and Loewenstein 2012, and Behnk et al. 2017), which suggests an additional channel through preferences: individuals in groups may assign a lower value to the moral outcome (probably through feeling less responsible for a bad outcome). To study the model's internal validity, I first argue that including a parameter to capture this factor would not meaningfully change the model's predictions. I then perform a lab experiment and show that subjects' behavior in a controlled setting is indeed well predicted by the model.

A second limitation is that the model is restricted to a one-shot simultaneous game without communication. While this structure is clearly stylized, I find evidence for the external validity of the model by comparing its predictions to the data in the meta-analysis by Fischer et al. (2011) on helping behavior and group size. Hence, despite the stylized structure of the model, I show that it captures a key element that determines whether one individual or a group is more likely to produce a moral outcome: the proportion of moral individuals in the situation.

References

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

Babcock, L., Recalde, M. P., Vesterlund, L., and Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3):714–747.

Baland, J.-M. and Platteau, J.-P. (1996). *Halting degradation of natural resources: is there a role for rural communities?* Food & Agriculture Org.

Barron, G. and Yechiam, E. (2002). Private e-mail requests and the diffusion of responsibility. *Computers in Human Behavior*, 18(5):507–520.

Bartling, B. and Özdemir, Y. (2017). The limits to moral erosion in markets: Social norms and the replacement excuse. *CESifo Working Paper Series*.

Behnk, S., Hao, L., and Reuben, E. (2017). Partners in crime: Diffusion of responsibility in antisocial behaviors. *IZA Discussion Paper*.

Bergstrom, T. (2017). The good samaritan and traffic on the road to jericho. *American Economic Journal: Microeconomics*, 9(2):33–53.

Bergstrom, T., Garratt, R., and Leo, G. (2019). Let me, or let george? motives of competing altruists. *Games and Economic Behavior*, 118:269–283.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, 102(5):841–877.

Bliss, C. and Nalebuff, B. (1984). Dragon-slaying and ballroom dancing: The private supply of a public good. *Journal of Public Economics*, 25(1-2):1–12.

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.

Bornstein, G. and Yaniv, I. (1998). Individual and group behavior in the ultimatum game: are groups more “rational” players? *Experimental Economics*, 1(1):101–108.

Brennan, G. and Lomasky, L. (1997). *Democracy and decision: The pure theory of electoral preference*. Cambridge University Press.

Campos-Mercade, P. (2018a). Helping behavior and group size. *AEA RCT Registry. June 17. https://www.socialscienceregistry.org/trials/2982/history/30845.*

Campos-Mercade, P. (2018b). Helping behavior and group size: The volunteer's dilemma explains the bystander effect. *Working paper*.

Chamberlin, J. (1974). Provision of collective goods as a function of group size. *American political science review*, 68(2):707–716.

Charness, G., Rigotti, L., and Rustichini, A. (2007). Individual behavior and group membership. *American Economic Review*, 97(4):1340–1352.

Charness, G. and Sutter, M. (2012). Groups make better self-interested decisions. *Journal of Economic Perspectives*, 26(3):157–76.

Chekroun, P. and Brauer, M. (2002). The bystander effect and social control behavior: The effect of the presence of others on people's reactions to norm violations. *European Journal of Social Psychology*, 32(6):853–867.

Clark, R. D. and Word, L. E. (1974). Where is the apathetic bystander? situational characteristics of the emergency. *Journal of Personality and Social Psychology*, 29(3):279.

Cox, J. C. (2002). Trust, reciprocity, and other-regarding preferences: Groups vs. individuals and males vs. females. In *Experimental business research*, pages 331–350. Springer.

Cryder, C. E. and Loewenstein, G. (2012). Responsibility: The tie that binds. *Journal of Experimental Social Psychology*, 48(1):441–445.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Darley, J. M. and Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4, Pt.1):377–383.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution*, 29(4):605–610.

Diekmann, A. (1993). Cooperation in an asymmetric volunteer's dilemma game theory and experimental evidence. *International Journal of Game Theory*, 22(1):75–85.

Eckel, C. C. and Grossman, P. J. (2005). Managing diversity by creating team identity. *Journal of Economic Behavior & Organization*, 58(3):371–392.

Falk, A., Neuber, T., and Szech, N. (2020). Diffusion of being pivotal and immoral outcomes. *The Review of Economic Studies*.

Falk, A. and Szech, N. (2013). Morals and markets. *Science*, 340(6133):707–711.

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

Fischer, P., Greitemeyer, T., Pollozek, F., and Frey, D. (2006). The unresponsive bystander: Are bystanders more responsive in dangerous emergencies? *European journal of social psychology*, 36(2):267–278.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., and Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, 137(4):517–537.

Franzen, A. (1995). Group size and one-shot collective action. *Rationality and Society*, 7(2):183–200.

Fromell, H., Nosenzo, D., Owens, T., Tufano, F., et al. (2017). Are victims truly worse off in the presence of bystanders? revisiting the bystander effect. *Working paper.*

Gaertner, S. L. (1975). The role of racial attitudes in helping behavior. *The Journal of Social Psychology*, 97(1):95–101.

Goeree, J. K., Holt, C. A., and Smith, A. M. (2017). An experimental examination of the volunteer's dilemma. *Games and Economic Behavior*, 102:303–315.

Harari, H., Harari, O., and White, R. V. (1985). The reaction to rape by american male bystanders. *The Journal of social psychology*, 125(5):653–658.

Hegstad, C. D. and Wentling, R. M. (2004). The development and maintenance of exemplary formal mentoring programs in fortune 500 companies. *Human Resource Development Quarterly*, 15(4):421–448.

Hillenbrand, A., Werner, T., and Winter, F. (2020). Volunteering at the workplace under incomplete information: Teamsize does not matter. *MPI Collective Goods Discussion Paper*, (2020/4).

Hillenbrand, A. and Winter, F. (2018). Volunteering under population uncertainty. *Games and Economic Behavior*, 109:65–81.

Huck, S. and Konrad, K. A. (2005). Moral cost, commitment, and committee size. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, pages 575–588.

Isaac, R. M. and Walker, J. M. (1988). Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, 103(1):179–199.

Isaac, R. M., Walker, J. M., and Williams, A. W. (1994). Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of public Economics*, 54(1):1–36.

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? we lie! why? experimental evidence on a dishonesty shift in groups. *Management Science*, 64(9):3995–4008.

Latané, B. and Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, 89(2):308–324.

Luhan, W. J., Kocher, M. G., and Sutter, M. (2009). Group polarization in the team dictator game reconsidered. *Experimental Economics*, 12(1):26–41.

Maaser, N. and Stratmann, T. (2019). Moral cost in weighted committee decisions. *Working paper.*

Near, J. (1996). Whistle-blowing: Myth and reality. *Journal of Management*, 22(3):507–526.

Nosenzo, D., Quercia, S., and Sefton, M. (2015). Cooperation in small groups: the effect of group size. *Experimental Economics*, 18(1):4–14.

Oliver, P. E. and Marwell, G. (1988). The paradox of group size in collective action: A theory of the critical mass. ii. *American Sociological Review*, pages 1–8.

Olson, M. (1965). *The Logic of Collective Action.* Cambridge: Harvard University Press.

Otsubo, H. and Rapoport, A. (2008). Dynamic volunteer's dilemmas over a finite horizon. *Journal of Conflict Resolution*, 52(6):961–984.

Palfrey, T. R. and Rosenthal, H. (1984). Participation and the provision of discrete public goods: a strategic analysis. *Journal of public Economics*, 24(2):171–193.

Panchanathan, K., Frankenhuis, W. E., and Silk, J. B. (2013). The bystander effect in an n-person dictator game. *Organizational Behavior and Human Decision Processes*, 120(2):285–297.

Philpot, R., Liebst, L. S., Levine, M., Bernasco, W., and Lindegaard, M. R. (2020). Would i be helped? cross-national cctv footage shows that intervention is the norm in public conflicts. *American Psychologist*, 75(1):66.

Ross, A. S. and Braband, J. (1973). Effect of increased responsibility on bystander intervention: Ii. the cue value of a blind person. *Journal of Personality and Social Psychology*, 25(2):254.

Rothenhäusler, D., Schweizer, N., and Szech, N. (2018). Guilt in voting and public good games. *European Economic Review*, 101:664–681.

Shaffer, D. R., Rogle, M., and Hendrlck, C. (1975). Intervention in the library: The effect of increased responsibility on bystanders' willingness to prevent a theft. *Journal of Applied Social Psychology*, 5(4):303–319.

Sherman, R. and Willett, T. D. (1967). Potential entrants discourage entry. *Journal of Political Economy*, 75(4, Part 1):400–403.

Shotland, R. L. and Heinold, W. D. (1985). Bystander response to arterial bleeding: helping skills, the decision-making process, and differentiating the helping response. *Journal of personality and social psychology*, 49(2):347.

Smith, R. E., Smythe, L., and Lien, D. (1972). Inhibition of helping behavior by a similar or dissimilar nonreactive fellow bystander.

Staub, E. (1970). A child in distress: The influence of age and number of witnesses on children's attempts to help. *Journal of Personality and Social Psychology*, 14(2):130.

Van Den Bos, K., Müller, P. A., and Van Bussel, A. A. (2009). Helping to overcome intervention inertia in bystander's dilemmas: Behavioral disinhibition can improve the greater good. *Journal of Experimental Social Psychology*, 45(4):873–878.

Weesie, J. (1993). Asymmetry and timing in the volunteer's dilemma. *Journal of Conflict Resolution*, 37(3):569–590.

Yang, W., Liu, W., Viña, A., Tuanmu, M.-N., He, G., Dietz, T., and Liu, J. (2013). Nonlinear effects of group size on collective action and resource outcomes. *Proceedings of the National Academy of Sciences*, 110(27):10916–10921.

Zhang, X. M. and Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review*, 101(4):1601–15.

Online Appendix A

This section extends the model in the main text by allowing agents to not only have preferences for the moral outcome to be implemented, but also to have warm-glow preferences (feel good by making the moral decision) and preferences to conform to the norm (feel good by acting as they believe that others act). If the preferences to conform to the norm are not too strong (such that agents mainly care about making the decisions others make), I show that Hypothesis 3—the model's main hypothesis—also holds in this case.

Agents' utility function consists of five parameters. First, playing M always carries a cost $c \geq 0$, and playing $\neg M$ does not carry any cost. Second, if the moral outcome is implemented, each agent i receives a moral payoff $b_i \geq 0$. Third, the mere fact of playing M gives those agents who play it a warm-glow payoff of $w_i \geq 0$. Fourth, an agent i dislikes making a decision that others do not make (or, equivalently, agent i likes making the same decision that others make). I assume $\mu_\alpha \alpha_i \geq 0$ to represent the intensity of this aversion when an expected share $\mu_\alpha \in [0, 1]$ of the other $n - 1$ agents play $\neg M$ and agent i plays M, and $(1 - \mu_\alpha)\beta_i \geq 0$ to represent the intensity of this aversion when an expected share $\mu_\beta = 1 - \mu_\alpha \in [0, 1]$ of the other $n - 1$ agents plays M and agent i plays $\neg M$. Note that one can interpret α_i and β_i as the intensity of the desire to follow others' behavior, which may be different depending on whether such behavior is to play M or $\neg M$. While this is arguably an oversimplified way to model preferences for norm conformity, it serves to illustrate how such preferences interact with the model.

I assume player i's utility U_i to be linear in these five components. Following the model in the main text, I assume that there are two types of agents:

1. *Immoral agents.* Agents who are never willing to play M, meaning that $b_{im} - c + w_{im} < 0$ (they prefer to play $\neg M$ when $n = 1$) and $-c + w_{im} + \beta < 0$ (they prefer to play $\neg M$ even if $n > 1$ and all others play M). The subscript im stands for "immoral."

2. *Moral agents.* Agents who would be willing to play m if no one else would, meaning that $b - c + w - \alpha > 0$. Since the analysis will only focus on these types of agents, to ease notation I do not use any subscript.

Before the game starts, nature picks with probability $\gamma \in (0, 1]$ a moral agent i with preferences $\{b_i, w_i, \alpha_i, \beta_i\} = \{b, w, \alpha, \beta\}$ and with probability $1 - \gamma$ an immoral agent i with preferences $\{b_i, w_i, \alpha_i, \beta_i\} = \{b_{im}, w_{im}, \alpha_{im}, \beta_{im}\}$. Agents know their own preferences but they only know the probability with which other agents are moral or immoral.

As in the main text, I focus the analysis on the unique Bayesian symmetric equilibrium. Let $\sigma_m(n, \gamma)$ be the probability with which a moral agent plays M when she is in a group of n agents.

Proposition 3. *If $\alpha + \beta \leq (n - 1)(1 - \gamma\sigma^*)^{n-2}b$, there exists a $\gamma^* \in (0, 1]$ such that the moral outcome is equally likely to be implemented by one agent alone and by a group of agents (i.e., GBE $= 0$). If $\gamma > \gamma^*$ the moral outcome is (weakly) more likely to be implemented by a group of one agent (i.e., GBE > 0). If $\gamma < \gamma^*$, the moral outcome is more likely to be implemented by a group of multiple agents (i.e., GBE < 0).*[21]

Proof. The proof is structured as follows. First, I show that, as long as $\alpha + \beta$ is not too large (such that agents mainly care about doing what others do), $\frac{\partial \sigma_m(n,\gamma)}{\partial \gamma} \leq 0$. Second, I show that $\frac{GBE}{P(1,\gamma)}$ strictly increases in γ. Finally, I use both results to prove the proposition.

1. ***Showing that*** $\frac{\partial \sigma_m(n,\gamma)}{\partial \gamma} \leq 0$.

   Define $EU(s)$ as expected utility of making decision $s \in \{M, \neg M\}$. To simplify notation, denote $\sigma_m(n, \gamma) = \sigma^*$. One can distinguish two cases:

   (a) $EU(M) > EU(\neg M)$. If the expected utility of playing M is higher than the expected utility of playing $\neg M$, then $\sigma^* = 1$. Because $\sigma^* \in (0, 1]$, an increase in the share of moral agents in this case cannot increase σ^*, implying that $\frac{d\sigma^*}{d\gamma} \leq 0$.

   (b) $EU(M) = EU(\neg M)$. Since the equilibrium is symmetric, moral agents only randomize between playing M and playing $\neg M$ if they are indifferent between both options. Define $\sigma^* \in (0, 1)$ as the probability that each agent plays M in equilibrium. Then, for an arbitrary agent to be indifferent

$$EU(M) = EU(\neg M)$$

[21] Note that in the main text $\gamma^* \in (0, 1)$ and here $\gamma^* \in (0, 1]$. The case where $\gamma^* = 1$ corresponds to the one where w is so high that all moral agents always play M regardless of the others' strategies.

$$b - c + w - \alpha(1 - \gamma\sigma^*) = (1 - (1 - \gamma\sigma^*)^{n-1})b - \beta\gamma\sigma^*.$$

Note that this is an implicit function that cannot be solved for σ^*. To find $\frac{d\sigma^*}{d\gamma}$ define

$$Z \equiv b - c + w - \alpha(1 - \gamma\sigma^*) - (1 - (1 - \gamma\sigma^*)^{n-1})b + \beta\gamma\sigma^*$$

$$= -c + w - \alpha + (1 - \gamma\sigma^*)^{n-1}b + \gamma\sigma^*(\alpha + \beta) = 0.$$

Using the implicit function theorem,

$$\frac{d\sigma^*}{d\gamma} = -\frac{\frac{dZ}{d\gamma}}{\frac{dZ}{d\sigma}} = \frac{\sigma^*(\alpha + \beta - (n-1)(1 - \gamma\sigma^*)^{n-2}b)}{\gamma(-\alpha - \beta + (n-1)(1 - \gamma\sigma^*)^{n-2}b)}.$$

This derivative is always negative whenever $\alpha + \beta$ is not too large, or more precisely when

$$\alpha + \beta \leq (n-1)(1 - \gamma\sigma^*)^{n-2}b.$$

Intuitively, whenever α and β are very large, agents eventually mainly care about behaving as they expect the other agents to behave. Keeping this limit in mind, in what follows I will assume that $\alpha + \beta \leq (n-1)(1 - \gamma\sigma^*)^{n-2}b$ such that $\frac{d\sigma^*}{d\gamma} \leq 0$.

Recall that $P(n, \gamma)$ is the probability that the moral outcome is implemented in a group of $n \geq 1$ agents when the proportion of moral agents is γ. Recall also that $GBE = P(1, \gamma) - P(n, \gamma)$.

2. ***Showing that*** $\frac{d}{d\gamma}(\frac{GBE}{P(1,\gamma)}) > 0$.

   Since γ is the share of moral agents for whom $b - c + w - \alpha > 0$, $P(1, \gamma) = \gamma$. To find $P(n, \gamma)$, note that the probability that exactly k of all the agents are moral is $\frac{n}{k!(k-n)!}\gamma^k(1-\gamma)^{n-k}$. In this case, the probability that the moral outcome is implemented is $(1 - (1 - \sigma^*)^k)$, where $\sigma^* \in (0, 1]$ represents the probability that a given agent plays M. Therefore, the total probability that the moral outcome is implemented can be written as

$$P(n, \gamma) = \sum_{k=1}^{n} \frac{n!}{k!(k-n)!}\gamma^k(1 - \gamma)^{n-k}(1 - (1 - \sigma^*)^k)$$

$$= \sum_{k=1}^{n} \frac{n!}{k!(k-n)!}\gamma^k(1-\gamma)^{n-k} - \sum_{k=1}^{n} \frac{n!}{k!(k-n)!}\gamma^k(1-\gamma)^{n-k}(1-\sigma^*)^k$$

$$= 1-(1-\gamma+\gamma(1-\sigma^*))^n = 1-(1-\gamma\sigma^*)^n,$$

and

$$\frac{GBE}{P(1,\gamma)} = \frac{\gamma-(1-(1-\gamma\sigma^*)^n)}{\gamma} = 1-\frac{1-(1-\gamma\sigma^*)^n}{\gamma}.$$

By taking the derivative with respect to γ

$$\frac{d}{d\gamma}(\frac{GBE}{P(1,\gamma)}) = -\frac{\gamma n(1-\gamma\sigma^*)^{n-1}(\sigma^*+\gamma\frac{d\sigma^*}{d\gamma})-1+(1-\gamma\sigma^*)^n}{\gamma^2},$$

which is positive whenever

$$Y \equiv -\gamma n(1-\gamma\sigma^*)^{n-1}(\sigma^*+\gamma\frac{d\sigma^*}{d\gamma})-(1-\gamma\sigma^*)^n+1 \geq 0$$

Note that Y monotonically decreases in $\frac{d\sigma^*}{d\gamma}$. Since $\frac{d\sigma^*}{d\gamma} \leq 0$, this means that showing $Y \geq 0$ for $\frac{d\sigma^*}{d\gamma} = 0$ implies $Y > 0$ for $\frac{d\sigma^*}{d\gamma} < 0$ as well. Assume then $\frac{d\sigma^*}{d\gamma} = 0$, and thus it remains to show that

$$Y \equiv 1-\gamma\sigma^* n(1-\gamma\sigma^*)^{n-1}-(1-\gamma\sigma^*)^n \geq 0.$$

Define $t = \gamma\sigma^*$ and define $f(t)$ such that

$$f(t) \equiv 1-tn(1-t)^{n-1}-(1-t)^n.$$

Next, I will show that $f(t_{min}) \geq 0$, where $t_{min} \in [0,1]$ is the t that minimizes this function. To find the t^* critical points

$$\frac{df(t)}{dt} = -n(1-t^*)^{n-1}+t^*(n-1)n(1-t^*)^{n-2}+n(1-t^*)^{n-1}$$

$$= t^*(n-1)n(1-t^*)^{n-2} = 0,$$

which only holds for $t^* = 0$ (a minimum) and $t^* = 1$ (a maximum). Since $f(t_{min}) =$

$f(0) = 0$, I conclude that $\frac{d}{d\gamma}(\frac{GBE}{P(1,\gamma)}) > 0$ (which is strict because $\gamma \neq 0$ by assumption and $\sigma^* \neq 0$).

3. ***Using the previous results to prove the proposition.***

   Recall that $GBE = P(1,\gamma) - P(n,\gamma) = \gamma - (1-(1-\gamma\sigma^*)^n)$. I will show that $GBE < 0$ for low γ, $GBE \geq 0$ for high γ, and show that there exists a unique γ^* such that $GBE = 0$.

   Pick a γ sufficiently small such that for a moral agent the expected utility of playing M is higher than the expected utility of playing $\neg M$, or $b - c + w - \alpha(1-\gamma\sigma^*) > (1-(1-\gamma\sigma^*)^{n-1})b - \beta\gamma\sigma^*$. This γ is guaranteed to exist because of the assumption that $b - c + w - \alpha > 0$ (which is the limit of the previous inequality when $\gamma \to 0$). In this case, agents will play M with probability $\sigma^* = 1$. Since $\gamma < 1$, this implies that $GBE = \gamma + (1-\gamma)^n - 1 \leq (1-\gamma)^n - (1-\gamma) < 0$.

   Pick $\gamma = 1$. Then $GBE = 1 + (1-\sigma^*)^n - 1 = (1-\sigma^*)^n > 0$ if $\sigma^* < 1$. If $\sigma^* = 1$, note that $\gamma^* = 1$ and that this implies that there does not exist any $\gamma > \gamma^*$. In this case, w is sufficiently high that all moral agents play M regardless of others. These results imply that GBE is negative for low γ and (weakly) positive for high γ. Note that $sign(GBE) = sign(\frac{GBE}{P(1,\gamma)})$. Since $\frac{GBE}{P(1,\gamma)}$ strictly increases in γ, this implies that there exists a unique γ^* such that $\frac{GBE}{P(1,\gamma^*)} = 0$ and therefore $GBE = 0$. □

Online Appendix B

This section assumes that there is a continuum of types that differ in their valuation of the moral outcome, b_i, which I label *moral preference.* I assume that the cumulative distribution function of the agents' moral preference is S-shaped: strictly convex for low moral preferences and strictly concave for high moral preferences. Note that in this setup the share of immoral agents—those agents who play $\neg M$ even when they are alone—is the share of agents whose moral preference is lower than the cost of playing M. Thus, to change the share of moral agents one must change either the moral preference distribution function or the cost of playing M. For simplicity, I will study changes in the cost of playing M (although I conjecture that results are equivalent when changing the moral preference distribution function). Note that, given a distribution function for moral preferences, a low cost of playing M implies that there is a low share of immoral agents, and a high cost of playing M implies that there is a high share of immoral agents. In line with Hypothesis 3, I show that there exists a unique cost c^*—that corresponds to a determined share of moral agents γ^*—that makes the GBE equal to zero. I show that any cost that is lower than c^*—where the share of moral agents is thus higher than γ^*—implies a positive GBE, meaning that the moral outcome is more likely to be implemented in a group of one than in a group of $n > 1$ agents. Furthermore, any cost that is higher than c^*—where the share of moral agents is thus lower than γ^*—implies a negative GBE, meaning that the moral outcome is more likely to be implemented in a group of n than in a group of one agent.

Agent i's utility function is $U_i = b_i - c$ if she plays M, $U_i = b_i$ if another agent plays M, and $U_i = 0$ if neither she nor another agent plays M. The moral preference parameter b_i is drawn from a commonly known probability distribution function $f(b)$ with a cumulative distribution function $F(b)$ bounded in the interval $b \in [0, 1]$. The function $F(b)$ is a differentiable S-shaped density function which is strictly convex for $b \in [0, \delta)$ and strictly concave for $b \in [\delta, 1]$, where $\delta \in (0, 1)$.

Proposition 4. *There exists a $\gamma^* \in (0, 1)$ such that the moral outcome is equally likely to be implemented by one agent alone and by a group of n agents (i.e., $GBE = 0$). If $\gamma > \gamma^*$, the moral outcome is more likely to be implemented by one agent alone (i.e., $GBE > 0$). If*

$\gamma < \gamma^*$, the moral outcome is more likely to be implemented by a group of multiple agents (i.e., $GBE < 0$).

Proof. Since $F'(b) \geq 0$ and since $F(b)$ is either strictly convex or strictly concave, $F(b)$ is strictly increasing for all $b \in (0, 1)$. This implies that $F(b)$ is a bijective function where every b corresponds to a unique value for $F(b)$. Note that the share γ of moral agents is $\gamma = 1 - F(c)$. Since $F(c)$ is bijective, then every $\gamma \in (0, 1)$ corresponds to a unique $c \in (0, 1)$. This implies that the proof is complete by showing that there exists a c^*, which corresponds to a unique γ^*, such that $GBE = 0, GBE > 0$ for $c < c^*$ and $GBE = 0$ for $c > c^*$.

If agent i is in a group of one agent, she will play M whenever her type b_i is higher than c. This implies that, a priori, the probability that the moral outcome is implemented when only one agent can implement it is the probability that $b_i \geq c$, so the probability that the moral outcome is implemented is $P(1, \gamma) = 1 - F(c)$.

I look for a symmetric Bayesian Nash equilibrium to find $P(n, \gamma)$. To calculate when an agent plays M, denote by p_k the a priori probability that an agent $k \in \{1, \ldots, n\}$ plays M and by p_{-i} the probability that at least one other agent plays M. Without loss of generality, assume that agents play M if they are indifferent between playing M and playing $\neg M$. Since agents play M as long as the utility of playing M is (weakly) higher than the utility of playing $\neg M$, agent i plays M if and only if

$$b_i \geq c + p_{-i} b_i \equiv b^* \tag{1}$$

and plays $\neg M$ if and only if $b_i < b^*$. Note that $b^* > c$, implying that the proportion of agents who are willing to play M is lower when there are n agents than when there is one agent (i.e., the BE). Now, the probability that at least one agent plays M is the probability that at least one agent has a b_i such that $b_i \geq b^*$, hence $P(n, \gamma) = 1 - F(b^*)^n$.

Thus,

$$\begin{aligned} GBE &= P(1, \gamma) - P(n, \gamma) \\ &= 1 - F(c) - (1 - F(b^*)^n) = F(b^*)^n - F(c). \end{aligned} \tag{2}$$

To find b^*, recall that agent i plays M if and only if $b_i - c \geq p_{-i} b_i$. Since p_{-i} is the probability that at least one agent $j \neq i$ picks a $b_j > b^*$, this inequality can be rewritten as

$$b_i - c \geq (1 - F(b^*)^{n-1})b_i.$$

Note that agent i will be indifferent between playing M and playing $\neg M$ only if $b_i = b^*$. Hence,

$$b^* - c = (1 - F(b^*)^{n-1})b^*,$$

which simplifies to

$$F(b^*)^{n-1}b^* = c. \tag{3}$$

Dividing equation (2) by c and using (3) yields $\frac{GBE}{c} = \frac{F(b^*)}{b^*} - \frac{F(c)}{c}$. Define $G(b) \equiv \frac{F(b)}{b}$ for $b \in (0, 1]$ and note that

$$sign(GBE) = sign(G(b^*) - G(c)). \tag{4}$$

I now derive some properties of $G(b)$ and prove the proposition by using (4). Let μ be the unique point at which $F(b) = b$. Such a b exists since $F(b)$ is S-shaped.

Lemma 1. *G(b) has the following four properties:*

1. *$G(b)$ is continuous in $b \in (0, 1]$.*
2. *$G(b) < 1$ for $b < \mu; G(b) = 1$ for $b = \mu$ and $b = 1$; and $G(b) > 1$ for $b \in (\mu, 1)$.*
3. *Let $b_{max} \equiv \operatorname{argmax}_b G(b)$. Such b_{max} exists and is unique.*
4. *$G'(b) > 0$ for $b \in (0, b_{max})$ and $G'(b) < 0$ for $b \in (b_{max}, 1)$.*

Proof. Point 1 follows from the assumptions that $F(b)$ and b are continuous. Point 2 follows from the definition of $G(b)$ and μ, and since $F(b)$ is S-shaped.

To derive Point 3, note first that $b_{max} \in [\delta, 1]$, which are the values of b for which $F(b)$ is concave. To see that, note that any maximum of $G(b)$ must satisfy the condition that

$G'(b) = 0$. Since $G'(b) = \frac{F'(b)}{b} - \frac{F(b)}{b^2}$, this implies that in such maximum

$$F'(b)b = F(b). \tag{5}$$

To show that $b_{max} \in [\delta, 1]$, suppose to reach a contradiction that (5) holds for $b \in (0, \delta)$. Since $F(b)$ is strictly convex in this range, by definition of convexity, for any $\lambda \in (0, 1)$ and $x \neq y$ for $x, y \in (0, \delta)$,

$$F(\lambda x + (1 - \lambda)y) < \lambda F(x) + (1 - \lambda)F(y).$$

Dividing both sides by λ and reordering yields

$$F(x) > F(y) + \frac{F(y + \lambda(x - y)) - F(y)}{\lambda},$$

that, by taking the limit as $\lambda \to 0$, becomes

$$F(x) > F(y) + F'(y)(y - x). \tag{6}$$

Now pick $y = b$ and $x = \epsilon$. Then, (6) becomes $F(\epsilon) > F(b) + F'(b)(y - \epsilon)$ and, as $\epsilon \to 0$,

$$F(b) < F'(b)b, \tag{7}$$

which contradicts (5).

This implies that, if b_{max} exists, it exists when $b \in [\delta, 1]$. To show that b_{max} exists, notice that $G(\mu) = 1, G(1) = 1$, and $G(b) > 1$ for $b \in (\mu, 1)$. Since $G(b)$ is continuous, it follows that there exists at least one value of b such that $b = b_{max}$.

To show that b_{max} is unique, note that

$$G''(b) = \frac{F''(b)}{b} + 2\left(\frac{F(b)}{b^3} - \frac{F'(b)}{b^2}\right). \tag{8}$$

Introducing (5) into the b terms of (8) yields

$$G''\left(\frac{F(b)}{F'(b)}\right) = \frac{F''(b)F'(b)}{F(b)} + 2\left(\frac{F'(b)^3}{F(b)^2} - \frac{F'(b)^3}{F(b)^2}\right) = \frac{F''(b)F'(b)}{F(b)}, \tag{9}$$

which is negative since $F(b) > 0, F'(b) > 0$ and $F''(b) < 0$ for $b \in [\delta, 1]$. These conditions follow from strict concavity of $F(b)$ for $b \in [\delta, 1]$ and $F'(b) \geq 0$.

Since $G''(b)$ is negative when the FOC is satisfied, any critical point of $G(b)$ in $b \in [\delta, 1]$ is a maximum, implying that such maximum is unique. Point 4 follows from the previous point that $G'(b) > 0$ for $b \in [\delta, b_{max})$ and $G'(b) < 0$ for $b \in (b_{max}, 1)$. To show that $G'(b) > 0$ also for $b \in (0, \delta)$, note that $G'(b) = \frac{F'(b)}{b} - \frac{F(b)}{b^2}$, and that this expression is positive by (7). □

Lemma 1 provides all the properties needed from $G(b)$ to prove that there exists a unique c^* (with a corresponding γ^*) such that $GBE = 0, GBE > 0$ for $c < c^*$ and $GBE < 0$ for $c > c^*$.

1. **There exists a** c^* **such that** $GBE = 0$. To see this, recall that $sign(GBE) = sign(G(b^*) - G(c))$. Since $G(b)$ is continuous for $b \in (0, 1], b^* \in (0, 1]$ and $c \in (0, 1), G(b^*) - G(c)$ is also continuous. Hence, this property can be shown by proving that there exists a cost such that GBE is negative and a cost such that GBE is positive.

   First, pick a c' such that $c' < \mu$. Note that since $F'(b) > 0$, there exists a unique $b^{*'}$ that satisfies the condition $c' = b^{*'}F(b^{*'})^{n-1}$. Since $c' < \mu$, and by the properties derived for $G(b) : G(c') < 1, G'(b) > 0$ for $b \in (0, b_{max})$, and $G(b) \geq 1$ for $b \in [\mu, 1]$. Since $c' < b^*$, it follows that $G(c') < G(b^{*'})$, which implies that, when the cost is $c', GBE > 0$.

   Second, pick a c'' such that $c'' = \text{argmax}_b\, G(b)$ is satisfied. This c'' corresponds to a $b^{*''}$ such that $c'' = b^{*''}F(b^{*''})^{n-1}$. Since $G(b)$ has a unique maximum, and because $c'' \neq b^{*''}, G(c'') > G(b^{*''})$. This implies that, when the cost is $c'', GBE < 0$.

2. c^* **is unique and** $GBE > 0$ **for** $c < c^*$ **and** $GBE < 0$ **for** $c > c^*$.

   To see this, note that, for the interval $b \in (0, \mu), G'(b) > 0$ and $G(b)$ is injective. Note however that $G(b)$ is non-injective in the interval $b \in [\mu, 1]$ because $G(\mu) = 1, G(1) = 1,$

and $G(b) > 1$ for $b \in (\mu, 1)$. In particular, since $G(b)$ has a unique critical point, for every $b \in [\mu, 1]$ and $b \neq b_{max}$ there exists a unique $b' \in [\mu, 1]$ such that $b \neq b'$ and $G(b) = G(b')$. This implies that any c^* such that $GBE = 0$ can only exist when $c^* \in [\mu, b_{max})$ (note that $b_{max} > \mu$ since $G(\mu) = 1$ and $G(b_{max}) > 1$). To see this, note that there does not exist any b^* such that $G(b^*) = G(c)$ for $c \in (0, \mu)$ since $G(c)$ is injective in this interval. Furthermore, there does not exist any b^* such that $G(b^*) = G(c)$ for $c \in [b_{max}, 1]$ because $G'(c) < 0$ in this interval and $c < b^*$. Thus, $c^* \in [\mu, b_{max}]$.

To show that c^* is unique, I show that $GBE > 0$ for any $c' < c^*$. Define b^* such that it satisfies $c^* = b^* F(b^*)^{n-1}$. Pick a c' such that $c' < c^*$ with a corresponding $b^{*'}$. Since $G'(b) > 0$ for $b \in (0, b_{max})$, and since c^* belongs to this interval, $G(c') < G(c^*)$. Since $c' = b^{*'} F(b^{*'})^{n-1}$, and $F'(b^{*'}) > 0$, then $b^{*'} < b^*$. There exist then three possible cases:

(a) If $b^{*'} > b_{max}$, then, since $G'(b) < 0$ for $b \in (b_{max}, 1), G(b^{*'}) > G(b^*)$. This implies that $GBE = G(b^{*'}) - G(c') > G(b^*) - G(c^*) = 0$.

(b) If $b^{*'} = b_{max}$, then since $G(b)$ takes its maximum value at b_{max}, and since b_{max} is unique, $G(b^{*'}) > G(c')$ for any c'. Thus, $GBE = G(b^{*'}) - G(c') > 0$.

(c) If $b^{*'} < b_{max}$, since $b^{*'} > c'$ and $G^{'}(b) > 0$ for $b \in [0, b_{max}), G(b^{*'}) > G(c')$. Thus, $GBE = G(b^{*'}) - G(c^{'}) > 0$.

The previous items show that $GBE > 0$ for any $c' < c^*$. The proof showing that $GBE < 0$ for any $c' > c^*$ is analogous. □

Online Appendix C

The following table shows the way in which dictator subjects were assigned to different pool compositions, depending on how many dictators paid in Decision 1 (D1). For example, when 5 of the 12 dictators paid in D1, the moral pool consisted of 4 dictators who paid (and 2 who did not pay) and the immoral pool consisted of 1 dictator who paid (and 5 who did not pay). The right column displays the number of sessions in which that number of dictators paid in D1.

# dictators who paid in D1	# dictators who paid in D1 in the moral pool	# dictators who paid in D1 in the immoral pool	Number of sessions
0	0	0	0
1	1	0	0
2	1	1	0
3	2	1	1
4	3	1	1
5	4	1	2
6	4	2	5
7	5	2	4
8	6	2	3
9	6	3	5
10	6	4	1
11	6	5	0
12	6	6	0

Hence, across all sessions, the number of pools that were created of each composition were:

# of dictators who paid in D1	Total number of pools
0	0
1	9
2	7
3	6
4	3
5	8
6	9

Online Appendix D

The following pages include a copy of the AEA RCT Registry of the experiment (AEARCTR-0002982). In sum, the main pre-registered hypothesis is Hypothesis 2 and the secondary one is Hypothesis 3. Hypothesis 1 (replicating the BE) was not pre-registered since it is a necessary condition for Hypothesis 3.

Note: The registry was updated on August 2020 to copy the text that had been written (before the experiment) in the ”Intervention (Hidden)” section to the ”Intervention (Public)” section. This information can be contrasted with the AEA RCT Registry.

Helping Behavior and Group Size

LAST REGISTERED ON AUGUST 13, 2020

VIEW TRIAL HISTORY

Restricted Access Edit Trial

Pre-Trial

Trial Information

GENERAL INFORMATION

Title
Helping Behavior and Group Size

RCT ID
AEARCTR-0002982

Initial registration date
May 12, 2018

Last updated
August 13, 2020 11:20 AM EDT

LOCATION(S)

Country
Germany

Region
Hamburg

PRIMARY INVESTIGATOR

Name
Pol Campos-Mercade

Affiliation
Lund University

Email
campos.mercade@gmail.com

OTHER PRIMARY INVESTIGATOR(S)

ADDITIONAL TRIAL INFORMATION

Status
Completed

Start date
2018-06-18

End date
2018-06-29

Keywords
Welfare

Additional Keywords
Help, Bystander Effect, Situations.

JEL code(s)

Secondary IDs

Abstract

Will a person in need of help be more likely to be helped when there are one or several potential helpers? Dozens of experiments have led social psychologists to conclude that the answer to this question depends entirely on the situation. This project uses game theory to predict in what situations one potential helper is more likely to provide help than a group of several potential helpers, and in what situations the opposite is true. The theoretical model concludes that in situations where few potential helpers are willing to help, then help is more likely to be provided when many people can help. However, in situations where most potential helpers are willing to help, help is more likely to be provided when only one person can help. I test this model in a lab experiment.

External Link(s)

REGISTRATION CITATION

Citation

Campos-Mercade, Pol. 2020. "Helping Behavior and Group Size." AEA RCT Registry. August 13. https://doi.org/10.1257/rct.2982-5.1.

Former Citation

Campos-Mercade, Pol. 2020. "Helping Behavior and Group Size." AEA RCT Registry. August 13. http://www.socialscienceregistry.org/trials/2982/history/73907.

Sponsors & Partners

Experimental Details

INTERVENTIONS

Intervention(s)

To understand this part, read first the Experimental Design.

The main intervention is in the second stage of the experiment. Here, I exogenously manipulate the composition of the subjects' pool. Some subjects therefore end up in a pool in which most AP chose to Pay in Decision 1, and some subjects end up in a pool in which most AP chose to Not Pay in Decision 1 (this, of course, depends on how many subjects chose to Pay in Decision 1; if for example all of them chose to Pay, then the two pools will consist only of AP who chose to Pay in Decision 1). The PP, who only make hypothetical decisions (and are told so), are always placed in a pool in which most AP chose to Not Pay in Decision 1.

Main hypothesis: Subjects in groups of multiple AP are more likely to choose to Pay when they are in a group in which most people chose to Not Pay in Decision 1 than when they are in a group in which most people chose to Pay in Decision 1. This hypothesis is tested between-subject (second stage) and within-subject (third stage).

Second hypothesis: If most people in a group chose to Pay in Decision 1, then the PP is better off in a group of one than in a group of multiple AP. If most people in a group chose to Not Pay in Decision 1, then the PP is better off in a group of multiple than in a group of one AP. This hypothesis is tested within-subject (third stage).

Intervention Start Date

2018-06-18

Intervention End Date

2018-06-29

PRIMARY OUTCOMES

Primary Outcomes (end points)

Help_second_1, Help_second_2, Help_second_3, Help_second_23 (between-test), Nethelp_second_23 (between-test), Help_third (main within-test).

Primary Outcomes (explanation)

Help_second_1 = Takes value 1 if the subject chose to Pay in the second stage when he is in a group with 1 AP and 0 otherwise.
Help_second_2 = Takes value 1 if the subject chose to Pay in the second stage when he is in a group with 2 AP and 0 otherwise.
Help_second_3 = Takes value 1 if the subject chose to Pay in the second stage when he is in a group with 3 AP and 0 otherwise.
Help_second_23 = Help_second_2 + Help_second_3.
Nethelp_second_23 = Help_second_23 - Help_first_2 - Help_first_3 (see secondary outcomes for the definition of Help_first)
Help_third = Includes all the variables about the decisions whether to Pay in the third stage, depending on group size and depending on the composition of the other AP in the group.

SECONDARY OUTCOMES

Secondary Outcomes (end points)

Help_first_1, Help_first_2, Help_first_3, Beliefs

Secondary Outcomes (explanation)

Help_first_1 = Takes value 1 if the subject chose to Pay in the first stage when he is in a group with 1 AP and 0 otherwise.
Help_first_2 = Takes value 1 if the subject chose to Pay in the first stage when he is in a group with 2 AP and 0 otherwise.
Help_first_3 = Takes value 1 if the subject chose to Pay in the first stage when he is in a group with 3 AP and 0 otherwise.
Beliefs = Belief elicitation across all stages.

EXPERIMENTAL DESIGN

Experimental Design

There will be 378 subjects. Each session will consist of 18 subjects. Subjects are initially divided between Active Participants (AP) and Passive Participants (PP). The AP earnings depend on their decisions, while the PP earnings depend on the decisions of the AP. However, the PP do not know that they are PP until the end. They therefore make hypothetical decisions throughout the experiment.

The game: Subjects are placed in groups of one, two, or three AP and one PP. Within each group, the AP start with 10€ and the PP starts with 0€. The AP then play the following game: they can choose to Pay 3€ or to Not Pay. To Pay 3€ means to get 7€ instead of 10€. If at least one AP in the group chooses to Pay, the PP gets 5€. If none of the AP chooses to Pay, then the PP gets 0€.

The experiment has three stages. In the first stage, subjects play this game in groups of one, two, and three AP (in random order). The decision that subjects make when they are in the group of one AP is called "Decision 1".

In the second stage, subjects play the same game but this time they get information about what other AP in their group chose in Decision 1. More concretely, they are told that the other AP in their group have been randomly selected from a "pool" of 6 AP. They are then told how many of these 6 AP of the pool chose to Pay in Decision 1.

The third stage uses the strategy method to elicit subjects' decision whether to Pay depending on how many of the other AP in their pool chose to Pay in Decision 1.

Experimental Design Details

Randomization Method

Randomization done by computer. Note that randomization is endogenous: those who chose to (Not) Pay in Decision 1 are more likely to be in a pool in which most people chose to (Not) Pay. However, given that someone chose to (Not) Pay in Decision 1, that subject is effectively randomly assigned into either of the pools.

Randomization Unit

Individual randomization.

Was the treatment clustered?

No

EXPERIMENT CHARACTERISTICS

EXPERIMENT CHARACTERISTICS

Sample size: planned number of clusters

378 subjects.

Sample size: planned number of observations

378 observations.

Sample size (or number of clusters) by treatment arms

Out of the 378 observations, 252 will be AP and 126 will be PP. The AP will be equally divided between those who see a pool in which most AP chose to Pay in Decision 1 and those who see a pool in which most AP chose to Not Pay in Decision 1. All the 126 PP will see situations in which most AP in their pool chose to Not Pay. Therefore, I expect about 252 observations where AP are in a pool in which most AP chose to Not Pay in Decision 1 and 126 where AP are in a pool in which most AP chose to Pay in Decision 1.

Minimum detectable effect size for main outcomes (accounting for sample design and clustering)

Power analysis for the main between-subject hypothesis: the weak bystander effect (defined as the difference between the percentage of bystanders helping when alone and when in a group) increases as the percentage of bystanders who chose to Pay in Decision 1 increases. If, for example, subjects help with the same frequency when they are alone regardless of the pool composition, then this hypothesis means that subjects are more likely to choose to Pay when they are in a group in which most people chose to Not Pay in Decision 1 than when they are in a group in which most people chose to Pay in Decision 1. This hypothesis is tested between-subject (second stage) and within-subject (third stage). I performed the power analyses through simulations (the STATA code is available upon request). I assume that 60% of the subjects choose to Pay in Decision 1. I only analyze the decisions of those subjects when they are in groups of 2 and 3 AP. I assume that 55% (70%) of the subjects choose to Pay in the pool in which most AP chose to (Not) Pay in Decision 1. Therefore, I study an effect size of 15 percentage points. The test I perform is a Wilcoxon ranksum test. Power to find a significant effect at the 5% level for the between-subject test: 85.3%. Power to find a significant effect at the 5% level for the within-subject test: 99%.

Supporting Documents and Materials

IRB

Analysis Plan