Review of Economic Studies (2020) **87**, 2205–2229 doi:10.1093/restud/rdz064

Advance access publication 24 February 2020

Diffusion of Being Pivotal and Immoral Outcomes

A. FALK
briq – Institute on Behavior & Inequality and University of Bonn
T. NEUBER
University of Bonn
and
N. SZECH
Karlsruhe Institute of Technology

First version received October 2014*; Editorial decision November* 2019*; Accepted January* 2020 (*Eds.*)

We study how the diffusion of being pivotal affects immoral outcomes. In our main experiment, subjects decide about agreeing to kill mice and receiving money versus objecting to the killing and foregoing the monetary amount. In a baseline condition, subjects decide individually about the life of one mouse. In the main treatment, subjects are organized into groups of eight and decide simultaneously. Eight mice are killed if at least one subject opts for killing. The fraction of subjects agreeing to kill is significantly higher in the main condition compared with the baseline condition. In a second experiment, we run the same baseline and main conditions but use a charity context and additionally study sequential decision-making. We replicate our finding from the mouse paradigm. We further show that the observed effects increase with experience, *i.e.*, when we repeat the experiment for a second time. For both experiments, we elicit beliefs about being pivotal, which we validate in a treatment with non-involved observers. We show that beliefs are a main driver of our results.

Key words: Diffusion of being pivotal, Group decisions, Morality, Replacement logic

JEL Codes: C91, C92, D01, D02, D23, D63, D71

1. INTRODUCTION

This article studies how groups favour moral transgression in diffusing responsibility and notions of being pivotal. Intuitively, acting in groups provides an excuse for acting immorally simply because an individual may perceive himself as not or only partly responsible for an outcome. To investigate the consequences of group settings that diffuse being pivotal, we ran two sets of experiments, varying the choice environment and contrast environments where subjects are fully pivotal with contexts where being pivotal is diffused by an exogenous change in organizational design. In the latter, subjects are organized into groups and individual decisions are aggregated such that the individual can easily believe that his decision is unlikely to be pivotal. Organizing people into groups and implementing a decision rule that does not require the support of all

The editor in charge of this paper was Adam Szeidl.

members for immoral action enables a simple "replacement logic" (see Sobel, 2010), which denotes the procedural phenomenon whereby people can mutually excuse their immoral behaviour with individual powerlessness in the face of others' immoral behaviour.

In our main experiment, the paradigm involves the trade-off between life and money. Subjects decide between receiving money and agreeing to kill mice versus not receiving money and objecting to the killing.[1] Importantly, mice used in the experiment are so-called "surplus" mice, all of which would have been killed without our intervention (see Section 2). Subjects learn about this default in a post-experimental debriefing. The paradigm is informed by the widely held view that harming others in an unjustified and intentional way is considered immoral.[2] We contrast two treatments: the *Baseline* treatment implements a simple binary choice where subjects either receive €0 for saving a mouse (Option A) or €10 for killing the mouse (Option B). In Baseline, subjects are hence fully pivotal. This condition serves as a comparison benchmark for the main *Simultaneous* treatment, in which eight subjects simultaneously decide between Option A and Option B. As in Baseline, a subject receives no money for choosing Option A and €10 for choosing Option B, irrespective of the other subjects' choices. However, if at least one subject chooses Option B, eight mice are killed. Thus, if a subject believes that at least one other subject is likely to choose Option B, he may no longer consider himself as being pivotal. In line with our argument, we find that the fraction of subjects choosing Option B is significantly higher in Simultaneous than in Baseline, even though—upon being pivotal—killing causes the death of eight mice rather than one. Moreover, the likelihood that a subject chooses to kill mice decreases with his belief of being pivotal. At the aggregate level, all mice are killed in Simultaneous.

Our second choice paradigm involves the binary decision between receiving €10 for oneself or donating €15 to a charity that supports children suffering from cancer. We replicate the two treatments from the first experiment as closely as possible (*BaselineC* and *SimultaneousC*) and additionally investigate experience effects, *i.e.*, whether the observed effects become larger if subjects repeat the same experiment one more time. For completeness, we also study a dynamic setting of diffusion of responsibility that mimics SimultaneousC but in which decisions are made sequentially (*SequentialC*). The charity experiment replicates the main effect found for the mouse conditions. The share of subjects choosing the selfish Option B is significantly higher in the simultaneous condition than in the baseline treatment. Moreover, choosing a second time in BaselineC on average does not affect the likelihood of donations but—as expected—induces more selfish choices in both SimultaneousC and SequentialC. In the latter, we additionally find that previous history matters for behaviour. In particular, learning that Option B has already been chosen essentially eradicates the choice of Option A further down the line.

Perceptions of being pivotal are central to the mechanism under study and they critically hinge on beliefs about the behaviour of others. This is why, in both experiments, we elicit beliefs and confirm that choices are strongly associated with the perceived likelihood of being pivotal. Given the critical role of beliefs, we ran a further treatment with non-involved subjects. In this condition, subjects read the instructions of all three treatments implemented with the charity paradigm and were asked to predict the results from the experiment. These independently elicited beliefs of spectators corroborate our above-mentioned findings. In particular, we find that the beliefs of spectators are very similar to those of subjects making a decision.

Organizational contexts that generate replacement arguments are pervasive at various levels of social interaction. They range from state-organized violence and corrupt bureaucracies to cheating

1. The study was approved by the ethical committee of the University of Bonn.

2. See, *e.g.*, Gert (2012, Section 1) on "The Definition of Morality," The Stanford Encyclopedia of Philosophy: "In this descriptive sense, although avoiding and preventing harm is common to all, 'morality' can refer to codes of conduct of different societies with widely differing content, and still be used unambiguously."

in sports, morally dubious market transactions, and malpractice within corporations. We discuss a few examples below. Some examples are more closely related to our simultaneous condition, others to the sequential choice context. Most real-world examples, however, share features of both. In this sense, our experimental group treatments represent limiting cases, where subjects decide either in isolation and complete uncertainty about other individuals' actions (Simultaneous and SimultaneousC) or with perfect information about previous choices and the exact timing and order of moves (SequentialC).

A striking example that closely corresponds to our simultaneous conditions is the practice of firing squads, which comprise of a group of executioners rather than a single person. For all members, shooting entails the personal advantage of avoiding disciplinary measures, and "technologically" one person who shoots his gun is sufficient to bring about the killing. From an individual member's perspective, being pivotal is diffused, as many people shooting at the same time implies that the killing is likely to happen, regardless of whether a particular member fires his gun or not. Moreover, members of firing squads are often randomly issued a gun containing a blank cartridge, which additionally diffuses being pivotal: even if a member of the squad shoots his gun, he remains uncertain whether or not he can effectively cause the killing at all. Apparently, these features reduce feelings of responsibility and facilitate participating in executions.

Corruption is another example that closely resembles the simultaneous decision-making context. Suppose that a citizen wants to gain illegitimate access to a public permit and therefore intends to bribe an official. He may approach different officials, but he only needs to find one single official who accepts the bribe. Since any official taking the bribe would do so secretly, there is no way to credibly signal honesty. If a given official is sufficiently certain that at least one of his colleagues is corrupt, he may now feel tempted to accept the bribe himself. This logic can give rise to an equilibrium where a large proportion of officials act corruptly. Doping in sports provides a similar example. Most athletes publicly state that they detest doping. However, many are later found guilty, with the road cyclist Lance Armstrong being an infamous example. This places athletes in a dilemma. They might generally object to cheating—at least because it jeopardizes the credibility of their discipline—but believe that others are doped anyway, which makes it seem more acceptable or even necessary to engage in doping themselves.

Reasoning about not being pivotal also helps to explain outcomes in markets that violate traders' own moral or fairness preferences. Here, a replacement argument prevails if traders prefer concluding a trade themselves over letting another trader perform the same transaction, even if trading creates unfair outcomes among the traders or imposes negative externalities on others. In cases where buying decisions create negative externalities, a frequent "excuse" is that "if I don't buy, another buyer will." On the opposite side of the market, suppliers of potentially harmful goods are in a similar situation, arguing that market demand would be met with or without their involvement. British Secretary of State Boris Johnson invoked an argument along these lines in October 2016 after allegations about weapons exported to Saudi Arabia being used for war crimes in Yemen. Faced with a motion in the House of Commons to suspend sales, he retorted that the respective members of parliament should "be in no doubt that we would be vacating a space that would rapidly be filled by other Western countries who would happily supply arms with nothing like the same compunctions or criteria or respect for humanitarian law" (Peck, 2016).[3]

The replacement logic also contributes to corporate crime. For example, Andrew Fastow—chief financial officer (CFO) of Enron from 1998 until 2001, who played a central role in concealing massive losses before the firm's bankruptcy in 2001 and served a six-year sentence

3. This is a refined version of the discussed argument in pointing at positive "side effects" associated with the United Kingdom taking an active role (see Glover and Scott-Taggart, 1975, pp. 177). Yet, the latter might often represent mere excuses rather than sound justifications.

in prison—himself drew the following parallel: "But the reality is, if at any point in my career I said 'time out, this is bullshit, I can't do it'…they would have just found another CFO, but that doesn't excuse it. It would be like saying it's OK to murder someone because if I didn't do it someone else would have" (Soltes, 2016, p. 255). The above quote underscores our main hypothesis, while it also highlights that behaviour in response to uncertainty about being pivotal may nevertheless be perceived as morally repulsive.[4] Note that the replacement logic draws on consequentialist moral thinking. By contrast, deontological moral reasoning would dictate doing the "right" thing regardless of being pivotal or not. The extent to which groups are vulnerable to transgression therefore crucially depends on the share of individuals following consequentialist versus deontological moral reasoning, respectively. We discuss the relative shares in the context of our experiments in Sections 3 and 4.

Our article is related to work on contextual factors affecting fair outcomes in the context of simple dictator, bargaining, or allocation games. While we focus on the role of beliefs about being pivotal, other mechanisms that have been identified to favour "unfair" outcomes are delegation or exploiting moral "wiggle rooms," as discussed, *e.g.*, in Bartling and Fischbacher (2012), Dana *et al.* (2007), Hamman *et al.* (2010), and Serra-Garcia and Szech (2018).[5] Falk and Szech (2013) analyse the malleability of moral outcomes in bilateral and multilateral market situations and Falk (2017) studies the role of status inequality.

The diffusion of being pivotal can be interpreted in terms of higher expected costs of acting morally because, in our group treatments, the probability of reaching a moral outcome when acting morally and foregoing the additional payment is smaller than one. In this sense, our findings are related to work by Andreoni and Miller (2002) and Fisman *et al.* (2007), who show that when exogenously varying the price of giving in simple dictator games, the observed willingness to share varies accordingly. Two important features differentiate our setup from this literature. First, we contrast a monetary benefit for oneself with a *moral* good. It remains to be shown whether people readily engage in trade-offs here as well. Second, we do not set the probability of being pivotal exogenously but it is determined endogenously by the behaviour of others, giving rise to equilibrium considerations.[6] Another related strand of literature in social psychology concerns the so-called bystander effect (see *e.g.* Latané and Darley, 1968 and for a recent overview Fischer *et al.*, 2011). Typical bystander experiments study helping behaviour in response to a staged emergency (*e.g.* the experimenter becomes injured). What sets our simultaneous treatments apart is that even if a subject opts for the moral outcome, he remains uncertain about whether the moral outcome is implemented or not, similar, *e.g.*, to firing squads. By contrast, in typical bystander experiments, this uncertainty does not exist. If a subject opts for helping, the person in need receives help. Furthermore, in a bystander experiment, while deliberating whether to help or not, subjects often observe that others do not help either. In our simultaneous-move setup, this type of social learning is ruled out. When deciding to kill a mouse or not to donate, respectively, subjects

4. Another example in this vein is the role of the replacement logic in the organization of the Holocaust (Arendt, 1963; Darley, 1992; Lifton, 1986). Lifton (1986) interviewed German doctors stationed in Auschwitz. They were operating in a nightmarish environment, with one of their objectives being to "select" prisoners who would be allowed to live while others would be immediately gassed. Being ordinary doctors, this activity was likely to be morally terrible and self-contradictory to them. Nevertheless, they engaged in the selection procedures. One of the frequently made justifications was that the "horrible machinery would go on," regardless of whether a particular doctor continued to participate. Replacement arguments suggesting the impossibility to stop ongoing moral crime were also used in the Nuremberg Trials as excuses for having participated in various kinds of atrocity under the Nazi Regime (see *e.g.* Crawford [2007] and references therein).

5. On the effects of institutions on values, see also Bowles (1998). On the role of authority, see Milgram [1974] (2009).

6. For an equilibrium analysis of group decisions in morally relevant contexts, see Rothenhäusler *et al.* (2018).

do not know whether other subjects also opt for the selfish option. The dynamic properties of observing others, however, are explicitly studied in our sequential treatment. Also, in a bystander experiment, participants need to realize that their help is required (and that it is better to step in than to hope that some other, say, more able helper will step in), while in our setup the consequences of decisions are straightforward. We also note that in our experiment, consequences are real, incentives are exactly specified, and the mechanism (beliefs about being pivotal) is explicitly measured.

The remainder of the article is organized as follows. In Section 2, we describe the design and implementation of the main experiment and develop our hypotheses. The results are presented in Section 3. Section 4 covers the charity experiment. We first present a replication of our main results and provide evidence for the validity of elicited beliefs. We then proceed by investigating an additional sequential condition. Finally, Section 5 concludes by summarizing the article and discussing additional observations.

2. EXPERIMENT

Avoiding and preventing unjustified harm is central to most notions of morality. It is this notion that informs the "mouse paradigm" used in our main experiment, which involves the trade-off between killing a mouse and receiving money versus saving a mouse life and receiving no money (Falk and Szech, 2013).[7] Subjects are explicitly informed that each mouse is a young and healthy mouse that will live for about two years if saved. For illustrative purposes, we present subjects the picture of a mouse on an instruction screen. We guarantee subjects that mice—if saved—live in an appropriate, enriched environment, jointly with a few other mice. Hence, in case subjects decided to save mice, these mice were kept alive in an enriched environment, with good feed and comfortable nesting material, precisely as stated in the instructions.

2.1. *Design*

Subjects are also informed in detail about the killing process. In the instructions (see Supplementary Appendix A), they read the following passage: "[T]he mouse is gassed. The gas flows slowly into the hermetically sealed cage. The gas leads to breathing arrest. At the point at which the mouse is not visibly breathing anymore, it remains in the cage for another 10 minutes. It will then be removed." To further rule out uncertainty about the decision context, subjects are shown a short demonstration video of the killing process. In the video, four mice first move vividly in the cage, then they successively slow down as more and more gas enters the cage. Eventually, they die, with their hearts visibly beating heavily and slowly.

It is important to stress that the mice used in the experiment were so-called "surplus" mice: these mice were bred for animal experiments but proved to be unsuited for scientific research. They were perfectly healthy, but keeping them alive would have been costly. It is common practice in laboratories conducting animal experiments to gas such mice. Thus, as a consequence of our experiment, many mice that would have otherwise all died were saved. Subjects were informed about this default in a post-experimental debriefing.[8]

7. Deckers *et al.* (2016) provide convergent and discriminatory validity of the mouse paradigm as a measure for morality. Killing is negatively related to agreeableness—one of the Big Five facets—which describes a tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others, and positively related to Machiavellianism, measuring a person's tendency to be unemotional and detached from conventional morality. Moreover, killing is not related to disposable income, whether students are professionally involved with animal research or animal experiments, or have a simple preference for animals, as expressed by having a pet at home.

8. While perceptions of the situation may have changed due to this information, the consequences were exactly the same and as stated in the instructions. In future research, it would be interesting to explore whether using an

Treatments We study the role of diffusion of being pivotal in contrasting two decision environments, one where subjects are fully pivotal (Baseline) and one where being pivotal is diffused by organizing subjects into groups (Simultaneous). The two decision contexts differ in terms of how likely it is that any given subject is pivotal, keeping overall moral and financial consequences identical. In Baseline, each subject decides about the life of one mouse. Subjects face a simple binary choice between Option A and Option B: Option A implies that the mouse will survive and that the subject receives no money, while Option B implies the killing of the mouse and receiving €10. The Baseline treatment informs us about the share of subjects who are willing to kill the mouse for €10 when obviously being pivotal.

In Simultaneous, subjects decide in groups of eight and are endowed with eight mice. As in Baseline, each subject faces an individual binary choice between Option A and Option B: Option A implies that a subject receives no money. If a subject chooses Option B, he receives €10. Individual monetary consequences are independent of other subjects' decisions. All subjects choose simultaneously. They know that if at least one subject chooses Option B, all eight mice are killed. Furthermore, they know that they will not receive feedback on whether the mice are ultimately killed or not (although it is obvious for a subject that the mice die if he chooses Option B). Note that we chose to endow a group with eight mice to keep the number of mice at the aggregate level identical to Baseline. Of course, the valuation of mice lives need not be proportional to the number of saved mice, but keeping numbers identical at the aggregate level allows for a clean comparison of the overall impact of group versus individual decision-making.

In Simultaneous, right after subjects have made their decision, we elicit beliefs about being pivotal. Subjects are asked to indicate the probability that all other seven group members have chosen Option A (*belief_pivotal*). We also ask subjects to estimate how many other subjects in their group have chosen Option B. They can enter any number from 0 to 7 and are paid €1 for a correct estimate (*belief_B*).

Procedure Two hundred and fifty-two subjects—mainly undergraduate university students from all majors—took part in the experiment, 124 subjects in Baseline and 128 in Simultaneous. Each subject participated only in one treatment condition. We used z-Tree as the experimental software (Fischbacher, 2007). Subjects were recruited using the software ORSEE (Greiner, 2004). At the beginning of an experimental session, participants received detailed information about the rules and the structure of the experiment. In all treatments, the experiment started only after all participants had answered several control questions correctly.

To reduce possible communication between subjects across sessions, the experiment was run on two consecutive days in six different rooms at the *Beethovenhalle*, the largest concert hall in Bonn. We set up six parallel, computerized labs in these rooms. Subjects received payments according to the rules of the experiment and an additional show-up fee of €20 to compensate for the remote location. In both treatments, subjects received their payments in a sealed envelope outside the room where the experiment had taken place. This way, neither other subjects nor the experimenters handing over the envelopes knew what a particular subject had earned. This procedure was explained in the instructions.

To ensure credibility, we stated right at the beginning that all statements made in the instructions were true—as is standard in economic experiments—and that all consequences of subjects' decisions would be implemented exactly as described in the instructions. We emphasized

alternative framing would affect decisions in response to institutional changes differently (compare evidence on the so-called omission–commission bias *e.g.* in Spranca *et al.*, 1991).

orally that the experimenters would personally guarantee the truthfulness of the instructions. Subjects were also invited to send us an email if they wanted to discuss the study.

2.2. *Hypotheses*

Our predictions start from the premise that most subjects follow consequentialist reasoning rather than deontological prescriptions. We expect that subjects in the Simultaneous treatment will engage in strategic considerations, thinking about how other subjects will decide. If they come to the conclusion that the likelihood of being pivotal is sufficiently small, subjects will find it justifiable to opt for the morally problematic Option B. Consequently, we would expect a higher share of subjects opting to kill in the group treatment compared with Baseline, in which subjects know that they are pivotal for certain.

To fix ideas, we normalize the utility from receiving €10 to one and the utility from receiving €0 to zero. There is a subjective moral cost $c_{n,i}$ for subject i associated with the death of $n=1$ or 8 mice, respectively. Furthermore, we denote by $belief\_pivotal_i \in (0,1]$ the subjective belief about the probability of being pivotal. If a subject chooses Option A and proves to be pivotal—*i.e.*, killing is averted—utility is given by 0. Otherwise, the resulting level of utility is $-c_{n,i}$. The subjectively expected utility from choosing Option A therefore amounts to $-(1-belief\_pivotal_i)c_{n,i}$. The utility from choosing Option B is always given by $1-c_{n,i}$. In making their decisions, deontological subjects disregard cost–benefit considerations and always choose Option A.[9] Any consequentialist subject chooses Option B if and only if the respective utility is at least as large as the subjectively expected utility from Option A or—equivalently—if $c_{n,i} \leq belief\_pivotal_i^{-1}$. Obviously, in the individual decision context, it holds for all subjects that $belief\_pivotal_i = 1$. By contrast, the belief about the chance of being pivotal in the simultaneous condition depends on beliefs about the behaviour of the other subjects in the same group.

This recursive relationship between subjects' decisions in Simultaneous can be understood as a strategic game between eight players whose types are characterized by their subjective moral costs $c_{8,i}$ and their respective moral conceptions, *i.e.*, whether they are deontologists or consequentialists. Types are independently drawn, with $d>0$ denoting the probability of a subject following deontological ethics and the distribution F of moral costs c_8 being continuous and having full support on the interval (a,b), with $a<1$ and $b>d^{-7}$. If we additionally impose that subjects hold correct beliefs given by $belief\_pivotal_i = p_i$, we can apply the concept of Bayesian equilibrium. According to the above discussion, individual behaviour follows a cut-off strategy in which an agent chooses Option B if $c_{8,i} \leq k_i$, with $k_i = p_i^{-1}$, and Option A otherwise. In our setup, a Bayesian equilibrium must feature strategies that are symmetric, *i.e.*, $k_i = k^\star$ for all agents. If any two agents within the same group used cut-off values that were different, the chance of being pivotal would be weakly higher for the agent whose cut-off value was higher. However, a weakly higher probability of being pivotal would imply that the cut-off value should be weakly lower, which is a contradiction.[10]

9. Alternatively, one could assume that deontologists take into account moral costs but always act as if they were deciding alone, *i.e.*, they deliberately abstain from equilibrium considerations. Indeed, Kant's categorical imperative requires people to "[a]ct only in accordance with that maxim through which you can at the same time will that it become a universal law" (cited from Kant, 1996, p. 73). Deontologists would then choose Option B if $c_{n,i} \leq 1$ and Option A otherwise (see also Roemer, 2010, 2015). The consequences for our analysis would be minor. For a discussion of the differences between consequentialist versus deontological reasoning, see Bénabou *et al.* (2018a,b).

10. Formally, assume that strategies are *not* symmetric. Players $1,\ldots,8$ form a group and—without loss of generality—it holds for their cut-off values that $k_1 < k_8$. For each agent, $p_i = \prod_{j\neq i}\{d+(1-d)[1-F(k_j)]\}$. It follows that $p_1 \leq p_8$ and thus $k_1 = p_1^{-1} \geq p_8^{-1} = k_8$, which gives the contradiction.

Consider a candidate k for an equilibrium cut-off value $k^\star$. In conjunction with the distribution of types, it implies a probability of being pivotal, which is given by $p(k)=\{d+(1-d)[1-F(k)]\}^7$. For an equilibrium cut-off value, a marginal subject for whom $c_{8,i}=k^\star$ needs to be indifferent between the two choice options. An equilibrium cut-off value is thus a fixed point for which $k^\star=p(k^\star)^{-1}$, *i.e.*,

$$k^\star=\left\{d+(1-d)\left[1-F\left(k^\star\right)\right]\right\}^{-7}. \tag{2.1}$$

The precise number and location of equilibria depends on the distribution of moral types. However, note that $p(k)^{-1}$ is not only strictly increasing in k but also continuous and its values range from 1 to d^{-7}. Thus, equilibrium cut-off values lie in the interval $\left[1,d^{-7}\right]$ and, by the intermediate value theorem, an equilibrium exists.[11]

As can be seen from equation (2.1), any equilibrium cut-off value $k^\star$ is always weakly larger than one, the latter being the cut-off under individual decision-making. In any equilibrium, the share of subjects choosing to kill is *strictly* larger than under individual decision-making as long as there exist any consequentialists ($d<1$), for whom we have assumed that some have moral costs smaller or equal than one ($F(1)>0$). Intuitively, some subjects choosing Option B even when fully pivotal reduce the likelihood of being pivotal for others, causing subjects with moral costs just above one to also choose Option B. Depending on the precise distribution of moral costs and the prevalence of deontologists, this leads other subjects with still higher moral costs to adjust their behaviour as well. In practice, the described moral unravelling will most likely reach an equilibrium only after some time and learning, similar to related experimental findings in, *e.g.*, market experiments where reaching an equilibrium typically requires several rounds of repetition. Even if an equilibrium has not been reached, however, the described moral unravelling suggests that the share of subjects choosing Option B should be higher in Simultaneous than in Baseline. This is our first hypothesis.

Hypothesis 1. *The share of subjects choosing Option B—thereby taking €10 and agreeing to kill—will be higher in Simultaneous than in Baseline.*

It is worth noting that as long as—for each individual—moral costs $c_{8,i}$ of killing eight mice are higher than moral costs $c_{1,i}$ of killing just one mouse, we tend to underestimate the role of being less pivotal in groups relative to Baseline. We could have endowed groups only with one mouse. In this case, we would expect even larger treatment effects. We opted for eight mice, however, to keep the maximum possible extent of harm fixed at the aggregate level when comparing treatments.

To the extent that an equilibrium has not been reached, subjects will most likely hold heterogeneous beliefs about the likelihood of being pivotal. We elicit these beliefs as part of our experimental procedure. According to the decision rule for consequentialists, the heterogeneity in beliefs should translate into corresponding differences in decisions, which is our second hypothesis.

Hypothesis 2. *In the Simultaneous treatment, the likelihood that a given subject opts for taking €10 and killing the mice decreases with the subjective probability assigned to being pivotal.*

In sum, the diffusion of being pivotal in groups leads consequentialist subjects to adjust their behaviour. The probability of being pivotal becomes small, making immoral behaviour more

11. Formally, in equilibrium it has to hold that $p(k^\star)^{-1}-k^\star=0$. Observe that $p(a)-a=1-a>0$ and $p(b)-b=d^{-7}-b<0$. Since the function $p(k)-k$ is continuous, it follows from the intermediate value theorem that an equilibrium exists.

attractive than when deciding individually. In addition, individual heterogeneity in the belief about the probability of being pivotal should translate into corresponding propensities to choose Option B. Hence, we expect that, on average, Option B is chosen more often in Simultaneous than in Baseline, and that—at the individual level—the likelihood of choosing Option B is inversely related to perceptions of being pivotal.

3. RESULTS

In presenting the results of our main experiment, we start with a treatment comparison. We then explicitly study the role of beliefs about being pivotal. According to our model, subjective beliefs along with observed choices imply bounds for each subject's individual moral costs. We use the joint distribution of beliefs and choices to estimate the distribution of subjective moral costs in the population and the prevalence of deontologists. Finally, we explore the implications of our estimates for welfare as well as for the equilibrium to which behaviour should ultimately converge.

3.1. *Choices and beliefs*

Our main result from the mouse experiment is shown in Figure 1, where we compare the shares of subjects choosing to kill in Baseline and Simultaneous, respectively. In Baseline, 46.0% of subjects choose Option B. In Simultaneous, the respective share is 58.6%, implying a difference of about 27%. This difference is significant ($p = 0.04$, two-sample test of proportions, two-sided) and confirms Hypothesis 1. At the aggregate level, the group impact is striking. While 46% of mice are killed in Baseline, *all* mice are killed in *all* groups in Simultaneous.

We have argued above that individual perceptions of being pivotal are critical in driving the increase in selfish behaviour in Simultaneous. Accordingly, we should observe that an individual's willingness to choose Option B decreases with his belief of being pivotal. This is indeed what we find. Recall that we asked subjects about the probability that all other group members had chosen Option A (*belief_pivotal*). Figure 2 displays the fraction of subjects choosing Option B depending on this belief.

The four categories in Figure 2 are based on quartiles of the belief distribution with respective percentage intervals of $[0, 3.5]$, $(3.5, 10]$, $(10, 35]$, and $(35, 100]$. In line with Hypothesis 2, the figure shows a clear negative relation between subjective perceptions of being pivotal and the likelihood of choosing Option B (Spearman rank correlation: -0.54, $p < 0.001$).[12]

3.2. *Implied moral costs*

In light of our formal framework introduced in Section 2.2, the observed heterogeneity in subjective beliefs about being pivotal provides a chance to estimate the distribution of moral costs—within the relevant choice context and subject population. Suppose, *e.g.*, that a consequentialist subject assigns a chance of 50% to the event of being pivotal. If the subject chooses Option B, one can infer that moral costs $c_{8,i}$ are at most $belief\_pivotal_i^{-1} = 2$. Conversely, if the subject chooses Option A, moral costs must be larger than two. To draw inferences about the distribution in the population, we make the following assumption.

12. The values of *belief_pivotal*—which we use here—and those of the incentivized *belief_B* are strongly and significantly correlated (Spearman rank correlation: -0.63, $p < 0.001$). The relationship between *belief_B* and choice of Option B is shown in Appendix B.1 and confirms the results presented here.

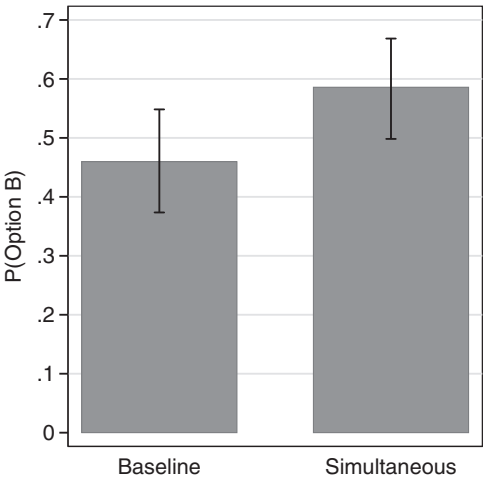

FIGURE 1

Treatment comparison.

Notes: Share of subjects choosing Option B in Baseline and Simultaneous. Error bars show 95% confidence intervals (based on logit transformations).

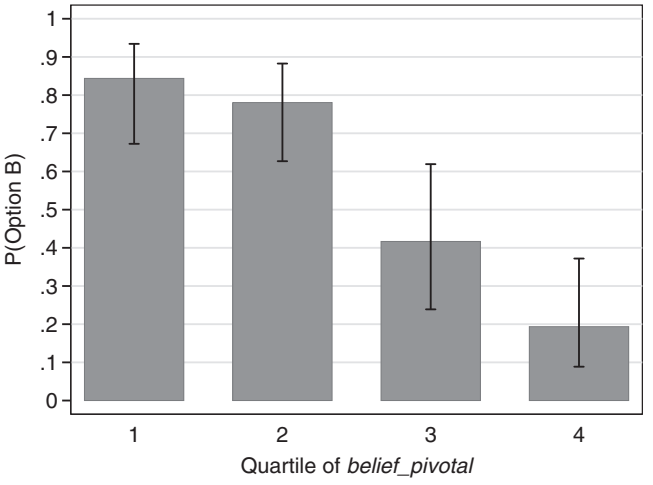

FIGURE 2

Belief quartiles (Simultaneous).

Notes: Share of subjects in Simultaneous choosing Option B depending on their belief of being pivotal. Error bars show 95% confidence intervals (based on logit transformations).

Assumption 1. *Moral costs are independent of the perceived likelihood of being pivotal.*

Then the share of consequentialist subjects choosing Option B among all those who believe that they are pivotal with a probability of 50% identifies the value of the distribution function F of moral costs at two. Similarly, the share of subjects choosing Option B among those who believe that they are pivotal with a probability of 25% identifies the value of the distribution function at four, and so on.

To be able to estimate the full distribution of moral costs, we need to impose some additional structure.

Assumption 2. *The subjective moral costs of consequentialist subjects follow a log-normal distribution F, with log-costs having mean μ and standard deviation σ.*

We can now write the probability of a given subject choosing Option B in terms of the cumulative distribution function of the standard normal distribution.

$$P(\text{Option B} \mid belief\_pivotal_i) = \begin{cases} \Phi\left(\dfrac{\ln\left(belief\_pivotal_i^{-1}\right)-\mu}{\sigma}\right) & \text{for consequentialists} \\ 0 & \text{for deontologists} \end{cases} \tag{3.2}$$

Next, consider a finite mixture model with two latent classes, one capturing consequentialists and the other deontologists. For consequentialists, a probit model is estimated that regresses the likelihood of choosing Option B on the log of the inverse probability of being pivotal and a constant. For deontologists, the probability of choosing Option B is always zero.

$$P(\text{Option B} \mid belief\_pivotal_i) = \begin{cases} \Phi\left[\beta_0+\beta_1 \ln\left(belief\_pivotal_i^{-1}\right)\right] & \text{for consequentialists} \\ 0 & \text{for deontologists} \end{cases}$$

In conjunction with equation (3.2), it follows that

$$\sigma = \beta_1^{-1} \qquad \text{and} \qquad \mu = -\frac{\beta_0}{\beta_1}.$$

We estimate this finite mixture model using the expectation–maximization (EM) algorithm, assigning subjects to latent classes in terms of probabilities.[13] The invariance property of maximum likelihood estimates then allows us to convert the point estimates for coefficients into estimates for the parameters of F, as described above.

Figure 3 visualizes the results. The left panel shows the density function f of moral costs c_8. The underlying estimates for the distributional parameters are $\hat{\mu} = 1.37$ and $\hat{\sigma} = 1.09$, corresponding to the mean and the standard deviation of log-costs, respectively. The expected value of moral costs is given by 7.098. We further estimate that the share of deontologists within our population of subjects is 13.6%, which is quite close to 17.9% of subjects choosing Option A despite being certain that they will not be pivotal.[14] The right panel uses these estimates to predict subjects' choices depending on their subjective beliefs about being pivotal for consequentialists (solid line) and deontologists (dashed line). Bubbles show observed choice probabilities by quartiles of the belief distribution, again separately for consequentialists and deontologists (solid and hollow, respectively). Deontologists never choose Option B. Consequentialists always choose Option B if they are certainly not pivotal, but this probability decreases to 10.3% if they believe that they are pivotal for sure. If subjects were deciding individually—as in Baseline—but about the lives of eight mice rather than just one, these estimates imply that 8.9% of them would choose Option B, which is much lower than the observed 46.0% opting to kill in Baseline.

13. If *belief_pivotal* is reported as 0%, we treat it as 0.1%.

14. Of course, it may also be the case that some subjects made mistakes. In this sense, deontologists comprise all people whose choice behaviour is unresponsive to beliefs about being pivotal.

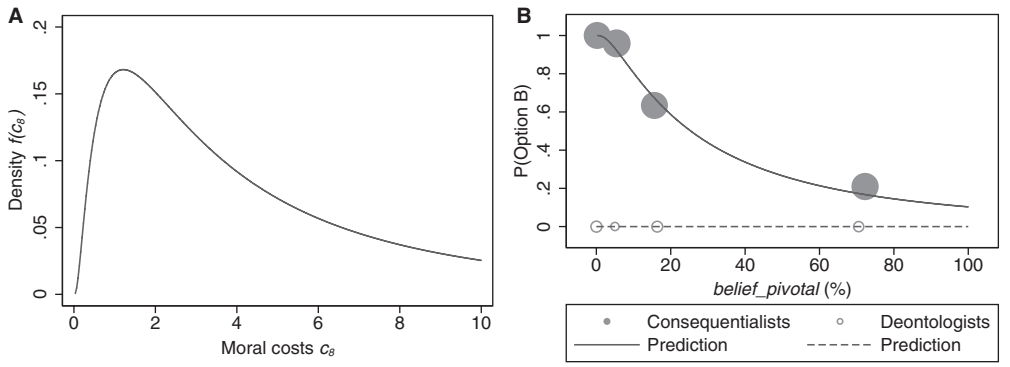

FIGURE 3

Moral costs (Simultaneous)

Notes: The left panel shows the estimated probability density function for moral costs c_8 of consequentialists in the Simultaneous treatment, denoted in multiples of the utility from receiving € 10. The right panel plots the implied probabilities of choosing Option B for different beliefs about the chance of being pivotal against observed shares in the experiment. The solid line depicts the predictions for consequentialists, which are given by $F(belief\_pivotal^{-1})$. Predictions for deontologists are given by zero and shown as the dashed line. For representing the data, subjects are first partitioned into quartiles of the belief distribution. Then, separately for consequentialists and deontologists within each quartile, probability-weighted average beliefs and shares of Option B are calculated. The sizes of bubbles correspond to estimates for the expected numbers of subjects. (A) Estimated moral costs and (B) observed versus predicted decisions.

3.3. *Welfare and equilibria*

To conduct a utilitarian welfare analysis based on average utility, we assume that the distribution of moral costs among deontologists is identical to the one for consequentialists, which can reasonably be interpreted as a lower bound. For ease of interpretation, we furthermore assume that utility is linear in money. Then, the average moral costs of killing eight mice across all subjects are equivalent to €70.98. Nonetheless, all mice are killed. All of those subjects who choose Option B secure a monetary payoff of €10, so that the average utility in Simultaneous (for observed behaviour) is equivalent to a loss of €65.12. If all subjects had chosen Option B, it would have been equivalent to a loss of €60.98. By contrast, if subjects were deciding alone, the utility would be weakly positive for everybody: all deontologists and those consequentialist subjects with moral costs above one (for eight mice) would choose Option A and receive a utility of zero, while consequentialists with moral costs between zero and one would opt for killing and receive utility corresponding to the subjective excess of utility from €10 over their cost of killing. The average level of utility would thus be equivalent to $(1-d)\int_0^1 (1-c)f(c)\mathrm{d}c \times €10$, which—according to our estimates—equals €0.31. Interestingly, a utility level of zero could also have been achieved in Simultaneous, had all subjects behaved as deontologists and saved the mice. This increased efficiency captures the intuition regarding why—from an evolutionary perspective—some degree of rule-based moral behaviour could indeed be expected (Alger and Weibull, 2013). However, our results point to a dominant role of consequentialist reasoning and question the relatively high fractions of Kantian types in survey data such as the trolley problem (Foot, 1967), where consequences are hypothetical rather than real.

Throughout this section, we have made use of the fact that beliefs about being pivotal are heterogeneous and generally large in comparison to actual outcomes. Both points are evidence that, in Simultaneous, no equilibrium has yet been reached. This is not surprising, given that it typically takes time and experience to arrive at correct beliefs. As a benchmark case, however, we want to conclude the analysis of the mouse experiment by analysing the predicted equilibrium, *i.e.*, the outcome we would eventually expect given sufficient experience and learning. As has been argued in Section 2.2, our experimental setup can generally feature multiple equilibria, which is

also true under the additional assumption of log-normally distributed moral costs. Intuitively, this is because choices in favour of Option A by different players act as strategic complements. Any given consequentialist player with moral costs greater than one will refrain from choosing Option B as long others choose the moral option with sufficient likelihood but will behave selfishly if only few others choose Option A. In any case, there must exist at least one (interior) equilibrium, since the existence of deontologists always assures a strictly positive likelihood of being pivotal, which is enough to make some consequentialists with very high moral costs choose Option A. The extent to which their effort to save mice is enough for making yet further consequentialists save the mice as well depends on the prevalence of such high-cost individuals. We can inspect concrete equilibria by plugging our estimates from Section 3.2 into the equilibrium condition given by equation (2.1). Appendix A provides a visualization, also including an analysis for a hypothetical smaller group size. We find that for the estimated distribution of moral types, the setting in Simultaneous has a unique equilibrium in which the share of consequentialists choosing Option A is virtually zero. Thus, the deterioration of moral behaviour in Simultaneous would have been even more pronounced if subjects had held rational beliefs. We would expect convergence to this equilibrium if subjects repeatedly faced decisions like in Simultaneous. In Section 4.1, we will find some indication that rational updating of beliefs indeed occurs in a similar setting and that behaviour changes accordingly.

4. REPLICATION AND EXTENSIONS

In this section, we employ a different setup. This second choice paradigm involves the binary decision between receiving €10 for oneself or donating €15 to a charity that supports children suffering from cancer. The charity treatments are essentially the same as in the mouse experiment, except that we use a different choice paradigm and study the role of experience as well as an additional sequential condition. As far as possible, we use the same design features, stake sizes (€10 for the selfish option), and wording and framing of choice options (the instructions are provided in Supplementary Appendix B). At the beginning of the experiment, subjects are made familiar with the charity, which is devoted to supporting children who suffer from cancer. In particular, the charity is engaged in psychological assistance and organizing leisure activities for children and their families, it helps with follow-up care and school-related issues, and supports parents and siblings as well as clinical research on cancer.

Charity treatments To check the replicability of our experimental results from the mouse paradigm, we study a baseline (BaselineC, "C" for "charity") and a simultaneous group condition (SimultaneousC), analogous to the mouse conditions. In BaselineC, subjects make the binary decision to either donate €15 (Option A) or keep €10 for themselves (Option B).[15] In SimultaneousC, subjects are in groups of eight and simultaneously choose either Option A or Option B, respectively. Choosing Option B implies receiving €10 and choosing Option A receiving no money, irrespective of the choices of other group members. A donation of €120 ($8 \times$ €15) for the charity is only initiated if all group members choose Option A. If one group member or more choose(s) Option B, the donation of €120 is destroyed. To study how a dynamic setting affects the diffusion of responsibility, we further run treatment SequentialC. This treatment is identical to SimultaneousC (including payments, donation, wording, etc.), except that subjects choose sequentially. It is randomly determined at which position a subject is asked to decide, one

15. Note that the design choice to donate €15 limits the plausibility of the argument that the €10 kept are spent on an alternative good cause.

subject being first, another second, up to position 8. Before making the binary decision (Option A or Option B), subjects are informed about their position (1 to 8) and the previous choice history, *i.e.*, how many subjects have previously chosen A and how many have opted for B. In both SimultaneousC and SequentialC, we also elicit beliefs analogous to Simultaneous in the mouse condition. Subjects are asked to indicate the probability that all other seven group members have chosen Option A. Responses are given in percent using a slider, with higher percentages reflecting a higher perceived likelihood of being pivotal for the respective subject (*belief_pivotal*).[16] We also ask subjects to estimate how many other subjects in their group have chosen Option B, with possible responses from 0 to 7 (*belief_B*). Correct answers are remunerated with €2.

To measure potential experience effects, all three conditions include a second round, which came to subjects as a surprise.[17] Subjects were told that they will make one more and final decision. In SimultaneousC and SequentialC, subjects learn whether at least one subject in their group has chosen Option B and thereby destroyed the donation and that they will make the same decision in the same group of eight, as in the first round. In SequentialC, they also know that they act in the same order, *i.e.*, each subject chooses at the same position as before. Payoffs and consequences are identical to the first round.

Charity procedures Four hundred and eighty-one subjects—mainly undergraduate university students from all majors—took part in the experiments, 121 subjects in BaselineC, 120 in SimultaneousC and 240 in SequentialC (30 groups). Each subject participated in only one treatment condition. We used oTree as experimental software (Chen *et al.*, 2016). Subjects were recruited using the software ORSEE (Greiner, 2004). At the beginning of an experimental session, participants received detailed information about the rules and structure of the experiment. In all treatments, the experiment only started after all participants had answered several control questions correctly. The experiments were run at the BonnEconLab in March 2017. Subjects received a show-up fee of €10.

4.1. *Replication and experience effects*

We begin by presenting the results for the two treatments that correspond to the ones in our main experiment. The main findings are summarized in Figure 4, which displays the share of subjects choosing Option B (not to donate) in conditions BaselineC and SimultaneousC, respectively. The dark bars show results from the first round, the light bars those of the second round (which was unexpected for subjects). Two observations can be made. First, we replicate the main result from the mouse experiment using a different choice paradigm. The share of subjects choosing Option B is significantly higher in SimultaneousC than in BaselineC, with means of 58.3% and 39.7%, respectively ($p = 0.004$, two-sample test of proportions, two-sided). The increase in selfish behaviour amounts to 47.0%, which is higher than the respective increase in the mouse condition. At the aggregate level, no single group in SimultaneousC effectively donated. Second, the detrimental effect of group decision-making on prosocial outcomes seems to increase with experience. Comparing the results between periods one and two reveals an increase in the likelihood of immoral choices upon learning the previous outcome of 12.5 percentage points

16. Beliefs are elicited in the same way in SimultaneousC and SequentialC, but we note that in the latter, beliefs will depend on position and responses are affected by previous play, *e.g.*, getting to know that Option B has already been chosen.

17. Of our 121 subjects who took part in BaselineC, only 79 took part in an experience condition, *i.e.*, in a second round. For the first two sessions (with 42 subjects) we only ran one round. In the analysis, we therefore either use 121 observations (Round 1) or 79 observations (Round 2), respectively.

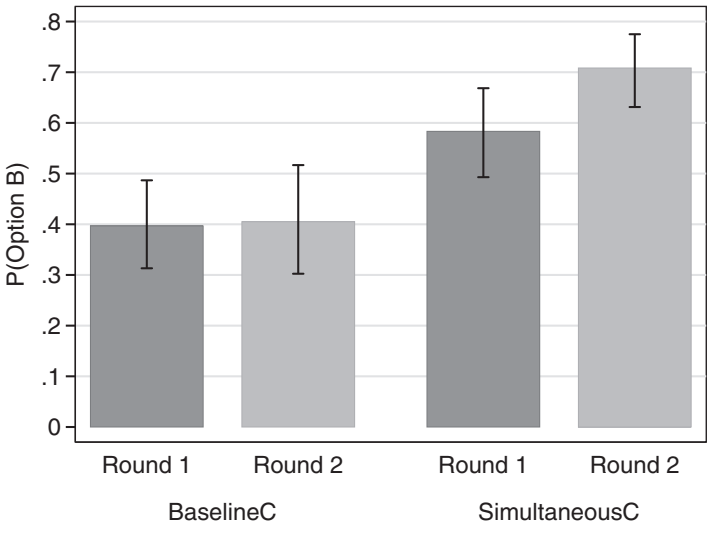

FIGURE 4

Comparison between BaselineC and SimultaneousC

Notes: Share of subjects choosing Option B in BaselineC and SimultaneousC, per round. Error bars show 95% confidence intervals (based on logit transformations), where standard errors are clustered at the group level for the second round of SimultaneousC.

($p = 0.03$, comparison of means, two-sided and with standard errors clustered at the group level for the second round of SimultaneousC). In sharp contrast, moral behaviour is not vulnerable to repetition in BaselineC, with an increase of Option B below one percentage point.

Analogous to the mouse experiment, we find that the association between the belief of being pivotal and choosing Option B is negative and statistically significant for SimultaneousC.[18] This relationship is shown in Figure 5, where we display the share of subjects choosing Option B depending on *belief_pivotal*. In SimultaneousC, among those who believe that they are not pivotal (estimated likelihood of being pivotal of 0%), 17.7% (three out of 17) of subjects choose Option A, presumably reflecting a Kantian kind of moral reasoning.

In both treatments, we observe some subjects switching from one choice option to the other between rounds. In the case of SimultaneousC, this switching is asymmetric, as reflected by the higher share of subjects choosing Option B in the second round. If beliefs about being pivotal are important drivers of behaviour, changes in beliefs should have predictive power for switching. In Table 1, we regress the choice in Round 2 on the choice in the first period and the change in the belief of being pivotal. There is a significant effect in the expected direction: subjects who consider themselves less pivotal in the second period than in Round 1 indeed become more likely to choose Option B in Round 2.

To summarize, we replicate the main findings from the mouse condition. Subjects are less likely to choose the morally desired action in SimultaneousC than in BaselineC (pertaining to Hypothesis 1 from Section 2.2) and beliefs about being pivotal seem to be critical (pertaining to Hypothesis 2). In addition, we document that selfish outcomes in groups tend to increase with experience in contrast to individual decisions, further supporting the crucial role of beliefs about being pivotal.

18. Again, both types of beliefs (*belief_B* and *belief_pivotal*) are significantly correlated (Spearman rank correlation: -0.35, $p < 0.001$). For results concerning the relationship between *belief_B* and choice of Option B, see Appendix B.1.

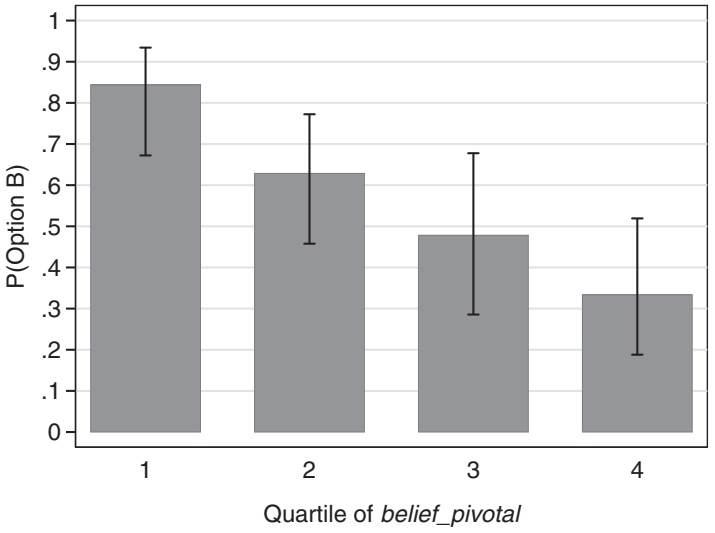

FIGURE 5
Belief quartiles (SimultaneousC)

Notes: Share of subjects choosing Option B in the first round of SimultaneousC depending on the belief of being pivotal. Error bars show 95% confidence intervals (based on logit transformations).

TABLE 1
Switching behaviour

	Dependent variable: Option B in Round 2		
	OLS (1)	Probit (2)	Logit (3)
Option B in Round 1	0.324***	0.333***	0.322***
	(0.0989)	(0.0962)	(0.0973)
Decrease in *belief_pivotal*	0.00398**	0.00385**	0.00389**
	(0.00169)	(0.00162)	(0.00175)
Constant	0.498***		
	(0.0880)		
Observations	120	120	120
Clusters	15	15	15
R^2	0.131		

Notes: Columns 2 and 3 report average marginal effects and average discrete changes due the binary choice in Round 1. Standard errors are clustered at the group level. $^{**}p < 0.05$; $^{***}p < 0.01$.

4.2. *Belief experiment*

A possible concern in interpreting beliefs is the potential endogeneity of beliefs due to motivated reasoning (Epley and Gilovich, 2016; Gino *et al.*, 2016). To limit the problem, we incentivized beliefs about the number of other participants choosing Option B in the mouse and charity treatments, such that subjects could earn additional money for accurate estimates. However, we also ran an additional belief experiment with non-involved observers. In the belief experiment, participants read the original instructions of treatments in the charity experiment (avoiding textual redundancies). We then ask them for the probability that a subject is in a group in which all other seven group members choose Option A (*belief_pivotal*). If the percentage answer (*belief_pivotal*) is correct within an interval of ±5 percentage points, they receive €2. Eighty-seven subjects

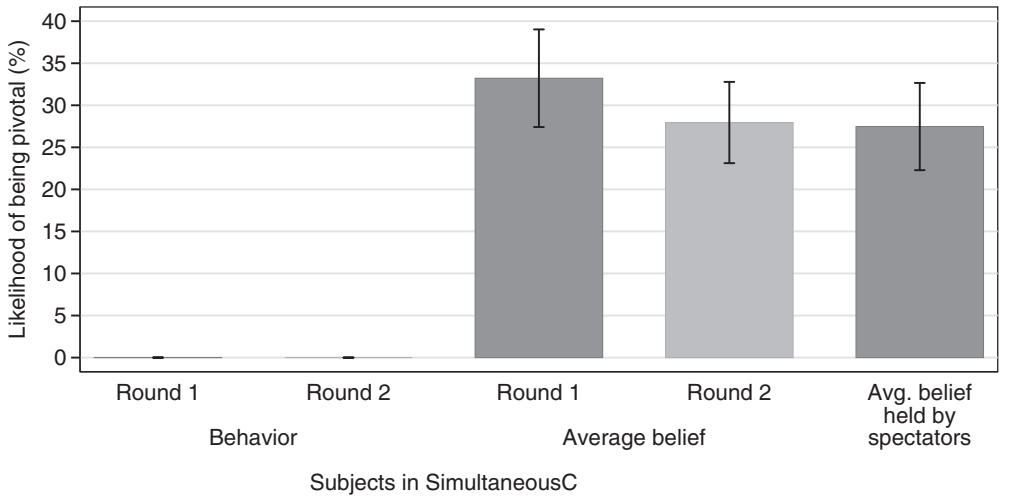

FIGURE 6

Belief comparison (*belief_pivotal*)

Notes: Likelihood of being pivotal, *i.e.*, the probability that all other seven members of a given subject's group choose Option A (in percent). Error bars show 95% confidence intervals, where standard errors are clustered for the second round.

participated in this condition, which was programmed with oTree (Chen *et al.*, 2016) and run at the BonnEconLab in March 2017.[19]

Figure 6 shows the results.[20] The actual probability for a subject to be in a group with all other seven group members choosing Option A was 0%, both in Rounds 1 and 2. In no single group, there were more than six subjects choosing Option A. A different way to estimate the actual probability of being pivotal is to use the whole distribution of choices and to calculate the likelihood—given the probability for Option A (41.7%)—of randomly being matched with seven group members who all choose Option A, which is 0.22%. This value is shown in the first bar and the analogous value of 0.02% for Round 2 in the second bar (the probability of Option A in the latter round is 29.2%). Bars 3 and 4 show subjects' average beliefs for Rounds 1 and 2, respectively. It is obvious that subjects heavily overestimate how likely it is that they are pivotal. While the shown average beliefs hide a substantial amount of heterogeneity, almost all subjects perceive themselves as being pivotal with a higher likelihood than what is true. Moving from Round 1 to Round 2, subjects adjust in the correct direction but still heavily overestimate their impact. Importantly, however, average beliefs of the spectators are not significantly different from those of active subjects in the first round of SimultaneousC ($p = 0.59$, Mann–Whitney U test, two-sided). On average, active subjects' beliefs are even slightly higher, suggesting that self-serving belief distortions do not play a dominant role in our main conditions.

19. Another interesting extension would be to investigate how behaviour depends on different sources of being pivotal. There is evidence that endogenously determined probabilities resulting from choices of other group members ("social risk") can give rise to different behaviour than probabilities determined by a correspondingly calibrated random device (see *e.g.* Bohnet *et al.*, 2008). In particular, if subjects in our experiment cared about fairness in relation to their fellow group members, they would potentially have additional reasons to act selfishly in the respective treatments: they could either wish to equalize their own monetary payoffs with the ones of other group members who are selfish, or they could feel "betrayed" if others did not cooperate in implementing the moral outcome. By contrast, if the probability of being pivotal was exogenously determined, social motives should be less relevant.

20. For corresponding results regarding *belief_B*, see Appendix B.2.

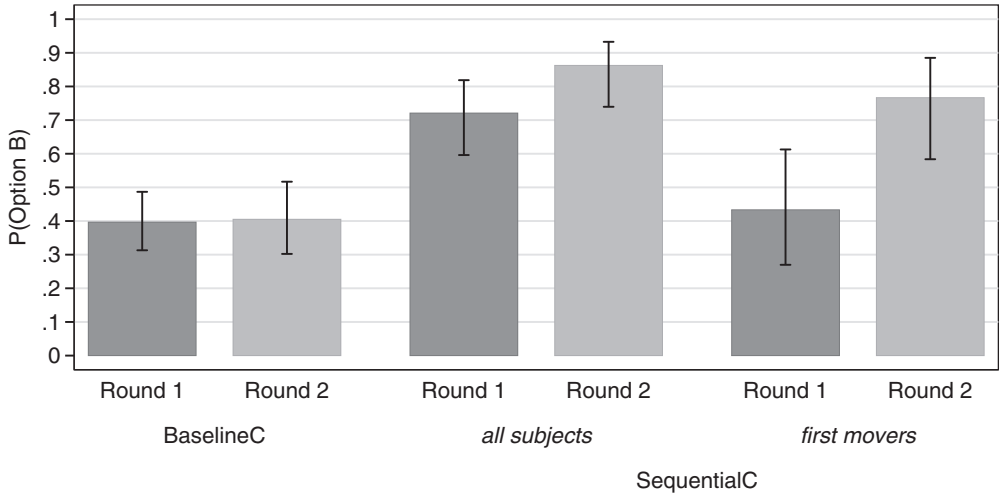

FIGURE 7

Comparison between BaselineC and SequentialC

Notes: Share choosing Option B among subjects in BaselineC, all subjects in SequentialC, and subjects in SequentialC who decide first in their groups, per round. Error bars show 95% confidence intervals (based on logit transformations), where standard errors are clustered at the group level for both rounds of SequentialC.

4.3. *Sequential decision-making*

We now turn to the sequential decision-making setup SequentialC. The central findings for this treatment are summarized in Figure 7. The overall share of participants choosing Option B in the first round of SequentialC is 72.1%, an increase of 81.7% relative to BaselineC. The difference between the two treatments is statistically significant ($p < 0.001$, comparison of means, two-sided and with standard errors clustered at the group level). This share increases by another 14.2 percentage points towards the second round ($p = 0.06$, comparison of means, two-sided and with standard errors clustered at the group level).[21] At the aggregate level, in both rounds, only two out of the 30 groups in SequentialC do not destroy the donation of €120.

Acting in a chain renders the specific position within the decision process relevant. Subjects deciding first in their group are of particular interest since, in a certain sense, they are in a similar situation as subjects in SimultaneousC. They have no information about others' behaviour in the given round and the consequences of the moral choice Option A for them depend on the behaviour of seven other subjects. In the first round, 43.3% of first movers choose Option B. Interestingly, this share is not significantly larger than in BaselineC ($p = 0.71$, two-sample test of proportions, two-sided). One can think of several plausible mechanisms contributing to this finding. First, the chance of being pivotal indeed seems to be higher for first movers in SequentialC than for subjects in SimultaneousC. Of the 17 cases in the first round where first movers choose Option A, two result in an actual donation. In light of the simple logic employed in Section 2.2, this might even be expected. Conditional on the donation not having been destroyed yet, choosing Option A becomes increasingly attractive the further down the line that a given subject decides, because the donation has to "survive" fewer remaining decisions. Subjects deciding at earlier positions should anticipate this recovery of moral behaviour over positions, incentivizing

21. Again, the two types of beliefs (*belief_B* and *belief_pivotal*) are significantly correlated (Spearman rank correlation: -0.65, $p < 0.001$).

them to preserve the donation themselves. Second, first movers overestimate their chance of being pivotal. The two surviving donations out of 17 cases where first movers choose Option A correspond to a likelihood of 11.8%, but first movers on average believe that it is 31.3%.[22] This could hint at exaggerated optimism regarding the possibility of acting as a prosocial role model (Gächter *et al.*, 2012, 2013). Third, subjects in SequentialC who choose Option B first in their groups are strongly identified with destroying the donation even in a constellation where, in fact, the decision would not have altered the outcome. This is because subjects at positions 2 to 8 make state-contingent choices, such that their counterfactual behaviour remains unknown. In particular, a subject who has chosen Option B will almost certainly observe that all subjects deciding at later positions will do the same but will not know what they *would have done* otherwise. The last two points lose most of their power in the second round. The large majority of first movers who had chosen Option A in the first round will learn that the donation was destroyed, meaning that they have not been pivotal. If they were hoping to be role models, they will feel frustrated. If they did not want to take the blame for choosing Option B first, they will now have a good excuse. Indeed, in the second round, the fraction of subjects who choose Option B increases to 76.7%, which is now significantly different from the second round of BaselineC ($p < 0.001$, two-sample test of proportions, two-sided). It thus seems that with experience, diffusion of being pivotal erodes moral behaviour also in the context with sequential decision-making.

Of course, the points discussed above may to some degree also apply to subjects deciding on other positions. It is therefore informative to consider the dynamics of choice behaviour in this treatment more broadly.

In Table 2, we explore the role of position and choice history in a simple panel regression framework using both rounds. In Columns 1 and 4, we regress a participant's choice of Option B on his position. Descriptively, subjects are more likely to choose Option B the further down the line that they decide. In Columns 2 and 5, we regress Option B on a dummy indicating that no other group member has chosen Option B yet ("not destroyed"). The respective intercepts are close to one and show that conditional on the donation already having been destroyed, almost all subjects choose Option B. The remaining subjects' decisions could reflect either a lack of attention or understanding (which is unlikely given the control questions and the prominent display of previous play on the decision screen) or a deontological notion of rule-based decision-making. More importantly, in the first as well as in the second round, subjects react strongly to being potentially pivotal, as reflected in the negative and significant coefficients (Columns 2 and 5, respectively). In Columns 3 and 6, we combine position and history and also include the interaction of the two. Turning to Round 1 (Column 3), the coefficients for the position as well as the interaction are insignificant, and the coefficient indicating that Option B has not yet been chosen is essentially identical to the one in Column 2. This suggests that subjects largely ignore their positions. Turning to the second round, this is no longer true. Now, the conditional probability of choosing Option B is generally high but decreases over positions. This can be interpreted as evidence of successful learning about moral types of subsequent subjects in the same group as well as about the imposed choice mechanism itself.

22. Note that for all 30 movers in the first round we also find that the belief of being pivotal and choice of Option A are significantly correlated in the expected direction (Spearman rank correlation: -0.68, $p < 0.001$ for *belief_pivotal* and 0.80, $p < 0.001$ for *belief_B*).

TABLE 2
Choice dynamics

	Dependent variable: Option B					
	Round 1			Round 2		
	(1)	(2)	(3)	(4)	(5)	(6)
Position (1–8)	0.0552***		−0.0121	0.0210*		0.00108
	(0.0139)		(0.0110)	(0.0115)		(0.00963)
Not destroyed		−0.626***	−0.607***		−0.458***	−0.154
		(0.0626)	(0.115)		(0.122)	(0.126)
Interaction			−0.0170			−0.126***
			(0.0278)			(0.0229)
Constant	0.473***	0.948***	1.013***	0.768***	0.968***	0.962***
	(0.0988)	(0.0228)	(0.0467)	(0.0798)	(0.0161)	(0.0572)
Observations	240	240	240	240	240	240
Clusters	30	30	30	30	30	30
R^2	0.0794	0.450	0.458	0.0196	0.313	0.435
Adj. R^2	0.0755	0.448	0.451	0.0155	0.310	0.428

Notes: OLS regression coefficient estimates, with binary choice option (Option B: destroy donation versus Option A: donate) as the dependent variable. Data come from the SequentialC treatment. *Position* is the position in the move order from 1–8, *Not destroyed* is a dummy that is 1 if all subjects in the respective group have chosen Option A thus far, and *Interaction* is the interaction of the two above variables. Standard errors in parentheses are clustered at the group level (30 groups). * $p<0.1$; *** $p<0.01$.

5. CONCLUSION

This article has documented the deterioration of moral outcomes in response to diffusion of being pivotal. Simple organizational changes from an individual decision context to group conditions increase moral transgression at the individual and even more so at the aggregate level.

In our main experiment, subjects decide to either kill mice in return for €10 or to save mice. In Baseline, subjects decide individually about the life of one mouse. In Simultaneous, subjects decide in groups of eight about the lives of eight mice. A single subject is enough to bring about the killing. We observe a statistically significant increase from 46.0% choosing to kill in Baseline to 58.6% in Simultaneous. In the group setting, all mice are killed. Our second paradigm closely resembles that of the first experiment but replaces killing mice with destroying charitable donations of €15 and €120 in the individual and group contexts, respectively. Analogously to the above comparison, we find a significant increase from 39.7% choosing the selfish option in BaselineC to 58.3% in SimultaneousC. To test for experience effects, we repeat the experiment in an unexpected second round. Repetition leaves the share in BaselineC virtually unchanged, while the share increases by another 12.5 percentage points in SimultaneousC. Using the charity paradigm, we also study a sequential context, in which eight subjects decide in a line and know whether the donation has already been destroyed. On average, 72.1% of subjects opt for destroying the donation and the share rises by another 14.2 percentage points towards the second round. Among subjects deciding first in their groups, 43.3% destroy the donation in the first round and 76.7% do so in the second round. Thus, with experience, immoral behaviour also deteriorates for first movers in the sequential choice context.

Consequentialism and deontological ethics have been centre stage in occidental moral philosophy for the last centuries. Empirical studies using the so-called trolley problem put forward by Philippa Foot (see also *e.g.* Greene *et al.* [2004] and Thomson [1976])[23] have provided

23. The quandary to be resolved in this problem is to follow either the deontologically warranted option (and not to throw a switch that will divert a trolley and kill one person) or the option preferred from a consequentialist perspective (killing the person to save five others).

support for the relevance of both. However, the evidence highlights the importance of situational and emotional factors. In contrast to the trolley evidence—which uses hypothetical outcomes—subjects in our experiment face real consequences. In all of our group treatments, we elicit beliefs about being pivotal. Subjects consistently respond to notions of being pivotal and only a few subjects appear to follow a Kantian conception. In Simultaneous, 17.9% of subjects who hold the belief that the chance of being pivotal is exactly zero choose Option A. In SimultaneousC, the respective share is almost identical with 17.7%. Finally, in SequentialC, of the 153 individuals for whom the group donation was already destroyed before, eight subjects (5.2%) nevertheless choose Option A. These numbers suggest the existence of deontological reasoning but they are quite low. Our findings thus question the relatively high fractions of Kantian types in survey data.

Using incentivized answers from non-involved observers, we show that there is no indication of subjects forming or reporting self-serving and thus biased beliefs in an attempt to justify selfish behaviour in our context. Generally, we find that beliefs about being pivotal are too high. Had they been more realistic, the willingness to engage in selfish behaviour may have been even more pronounced. In this sense, it is conceivable that repeated interactions with learning possibilities even further increase the likelihood of immoral outcomes, as we observe in the second round of our experiment using the charity paradigm. Overestimating one's sense of being pivotal could point to a human tendency to overestimate one's impact in general. This may well extend to other (non-moral) contexts and seems worth further investigating, *e.g.*, in voting contexts (Duffy and Tavits, 2008). In this context, Quattrone and Tversky (1984) argue and provide evidence that people use their own actions as prognostic for the behaviour of others, therefore trying to "induce" others to behave in a desired way even when no causal impact can exist. Another possible reason for overestimating one's impact could come from a desire for meaning, self-attribution and -determination, as well as for motivating action in general. Such a desire for self-efficacy is already known in the context of the so-called IKEA effect (*e.g.* Norton *et al.*, 2012).

While the focus of this article is to highlight possible negative consequences of organizational design on moral behaviour, the reverse inference is, of course, our main interest. Our findings suggest that organizations aiming to promote morality should reduce diffusion of being pivotal and instead attribute individual responsibility to their members.

Acknowledgments. We thank K. Albrecht, S. Altmann, R. Bénabou, T. Dohmen, J. Engel, D. Engelmann, S. Gächter, P. Heidhues, D. Huffman, S. Jäger, F. Kosse, B. Kőszegi, F. Krämer, G. Loewenstein, F. Rosar, J. Sobel, N. Schweizer, F. Zimmermann, and participants at various seminars for helpful comments. We thank M. Antony, T. Graeber, T. Wölk, and, in particular, S. Walter for excellent support. Falk acknowledges financial support by the German Research Foundation (DFG) through the Leibniz Program and by the European Research Council (ERC Advanced Grant 340950). The study was approved by the ethical committee of the University of Bonn (reference number: 066/12).

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

APPENDIX

A. EQUILIBRIA

We inspect the equilibrium condition developed in Section 2.2 using the parameter values for the share of deontologists d and for the log-normal distribution F of moral costs, μ and σ, estimated in Section 3.2 for individuals deciding over the lives of eight mice. To gain a better intuition, we generalize the condition for an equilibrium to exist at a cut-off value of $k^\star$ given by equation (2.1) to the case of a groups size of n.

$$k^\star = \left\{d + (1-d)\left[1 - F\left(k^\star\right)\right]\right\}^{1-n}$$

The two panels of Figure A1 provide a visual inspection of this equilibrium condition. Both are identically constructed but vary in their scale. Dashed lines show inverse probabilities of being pivotal as functions of the cut-off value k for

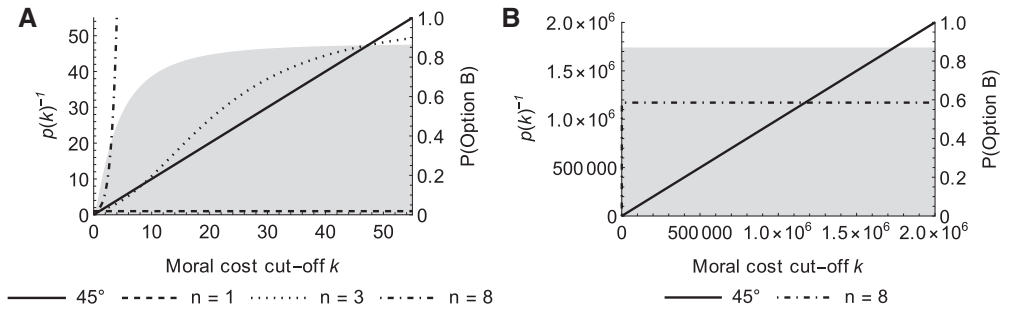

FIGURE A1

Equilibria (Simultaneous)

Notes: Numbers on the horizontal axes denote multiples of utility from receiving €10. The 45° line is drawn solidly. Dashed lines visualize the function $p(k)^{-1}$ for different (hypothetical) group sizes n. Values are given on the left axes. The shaded areas represent the cumulative distribution function $(1-d)F(k)$ of subjects choosing Option B. Values are given on the right axes. (A) Different (counterfactual) group sizes and (B) unique equilibrium for $n=8$.

moral costs at which subjects switch from choosing Option B to Option A. Equilibria are intersections of dashed lines with the solid 45° line. The left panel of Figure A1 shows that for $n=3$ (and still assuming the life of eight mice being at stake), there would exist three equilibria: one at 1.35 in which still only 14.0% of subjects would choose Option B, one at 8.46 in which 65.6% would choose Option B, and one at 47.35 in which 85.5% would do so. For our actual case of $n=8$, only a single equilibrium exists, which can be seen in the right panel of Figure A1. In this equilibrium, essentially all consequentialists choose Option B.

B. RESULTS FOR *BELIEF_B*

B.1. *Beliefs and choices*

We have established in Figures 2 and 5 (Sections 3.1 and 4.1, respectively) that beliefs about being pivotal are strongly associated with the propensity to choose Option B in both Simultaneous and SimultaneousC. We have used *belief_pivotal*, the percentage belief about the likelihood of being pivotal, because it is directly part of the formal analysis in Section 2.2. However, we also asked subjects about their belief regarding the number of other subjects in their group who chose Option B (*belief_B*), and the elicitation of this belief was incentivized. We show below that the same kind of relationship as for *belief_pivotal* can also be found for *belief_B*.

Figure B1A shows fractions of subjects choosing Option B depending on *belief_B* for subjects in Simultaneous. The categories are based on quartiles of the belief distribution and are given by the belief intervals $[0,2]$, $(2,4]$, $(4,6]$, and $(6,7]$. We see a monotonous increase in the propensity to choose Option B over belief quartiles, which is the expected mirror image of the Figure 2.[24] In particular, we see a strong increase between the first two quartiles, and the increases seem to fade out for the higher quartiles. In light of our framework, this is intuitive: subjects who believe that very few others—and thus potentially none—will choose Option B are highly reluctant to do so themselves, while for high expected numbers the precise beliefs do not matter a lot.

Figure B1B replicates the above relationship in the charity experiment, *i.e.*, for SimultaneousC. Again, we find a general increase of the share of subjects choosing Option B over quartiles, which correspond to intervals of $[0,2]$, $(2,4]$, $(4,5.5]$, and $(5.5,7]$. Again, the increase between the first two quartiles is pronounced, while the difference between the last two quartiles is insignificant. Thus, the analysis of the relationship between *belief_B*, an indirect measure for the belief of being pivotal, and choice of the immoral option lends additional support to Hypothesis 2.

B.2. *Belief experiment*

We report results from the belief experiment in which participants read the original instructions of treatments in the charity experiment (avoiding textual redundancies) and reported incentivized estimates corresponding to *belief_pivotal* and *belief_B* in SimultaneousC.

24. The Spearman rank correlation between *belief_B* and choice of Option B in Simultaneous is 0.65 ($p<0.001$).

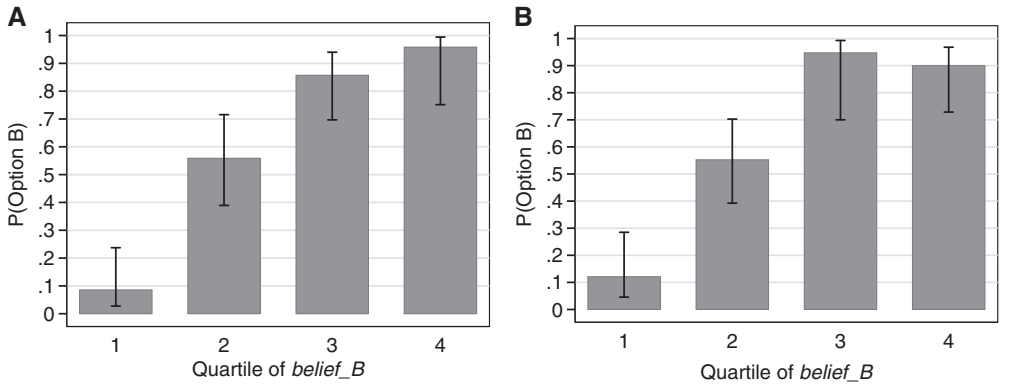

FIGURE B1

Belief quartiles for *belief_B* in Simultaneous and SimultaneousC (Round 1)

Notes: Share of subjects in the respective treatment choosing Option B depending on their belief about the number of other group members choosing Option B (*belief_B*). Error bars show 95% confidence intervals (based on logit transformations) (A) Simultaneous and (B) SimultaneousC (Round 1).

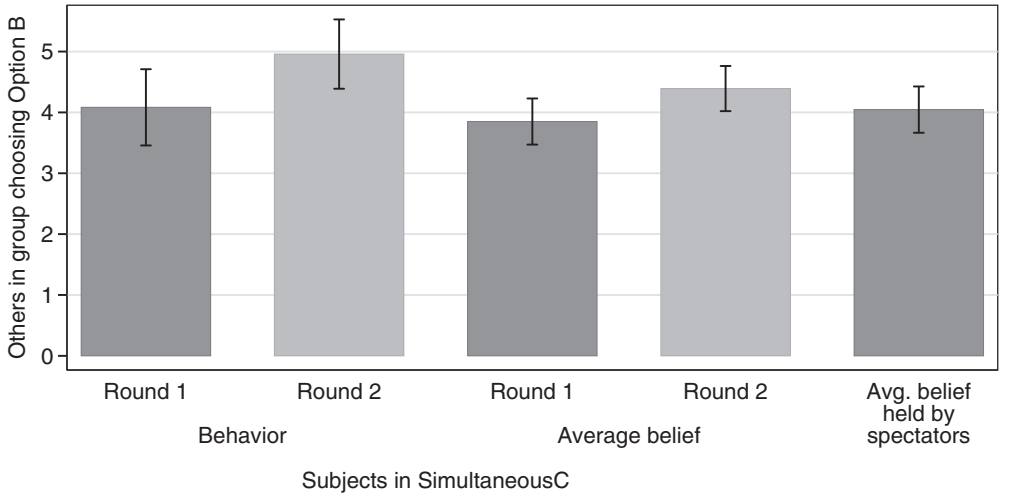

FIGURE B2

Belief comparison (*belief_B*)

Notes: Number of other group members choosing Option B (0–7). Error bars show 95% confidence intervals, where standard errors are clustered for the second round.

For *belief_B*, results are qualitatively quite similar to those concerning *belief_pivotal* (see Section 4.2). Figure B2 shows the actual behaviour of subjects in Rounds 1 and 2 (first two bars), average beliefs in Rounds 1 and 2 (bars 3 and 4), as well as average beliefs of spectators (fifth bar). The number of subjects choosing Option B increases from Round 1 to 2, which is reflected in changes in the beliefs of subjects. In contrast to *belief_pivotal*, however, subjects are overall much more accurate about actual outcomes.[25] Importantly, as for *belief_pivotal*, the average beliefs of active subjects and spectators are not statistically significantly different (comparison of bars 3 and 5 in Figure B2; $p = 0.59$, Mann–Whitney U test, two-sided).

25. A possible explanation is that subjects found estimating absolute numbers easier than estimating a probability.

REFERENCES

ALGER, I. and WEIBULL, J. W. (2013), "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching", *Econometrica*, **81**, 2269–2302.
ANDREONI, J. and MILLER, J. (2002), "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism", *Econometrica*, **70**, 737–753.
ARENDT, H. (1963), *Eichmann in Jerusalem* (New York, NY: Viking Press).
BARTLING, B. and FISCHBACHER, U. (2012), "Shifting the Blame: On Delegation and Responsibility", *Review of Economic Studies*, **79**, 67–87.
BÉNABOU, R., FALK, A. and TIROLE, J. (2018a), "Eliciting Moral Preferences" (Mimeo).
BÉNABOU, R., FALK, A. and TIROLE, J. (2018b), "Narratives, Imperatives, and Moral Reasoning" (NBER Working Paper 24798, National Bureau of Economic Research (NBER), Cambridge, MA).
BOHNET, I., GREIG, F., HERRMANN, B. ET AL. (2008), "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States", *American Economic Review*, **98**, 294–310.
BOWLES, S. (1998), "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions", *Journal of Economic Literature*, **36**, 75–111.
CHEN, D. L., SCHONGER, M. and WICKENS, C. (2016), "oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments", *Journal of Behavioral and Experimental Finance*, **9**, 88–97.
CRAWFORD, N. C. (2007), "Individual and Collective Moral Responsibility for Systemic Military Atrocity", *Journal of Political Philosophy*, **15**, 187–212.
DANA, J., WEBER, R. A. and KUANG J. X. (2007), "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness",. *Economic Theory,* **33**, 67–80.
DARLEY, J. M. (1992), "Social Organization for the Production of Evil", *Psychological Inquiry*, **3**, 199–218.
DECKERS, T., FALK, A., KOSSE, F. ET AL. (2016), "Homo Moralis: Personal Characteristics, Institutions, and Moral Decision-Making", (IZA Discussion Paper 9768, Institute for the Study of Labor (IZA), Bonn, Germany).
DUFFY, J. and TAVITS, M. (2008), "Beliefs and Voting Decisions: A Test of the Pivotal Voter Model", *American Journal of Political Science*, **52**, 603–618.
EPLEY, N. and GILOVICH, T. (2016), "The Mechanics of Motivated Reasoning", *Journal of Economic Perspectives* **30**, 133–140.
FALK, A. (2017), "Status Inequality, Moral Disengagement and Violence" (CESifo Working Paper 6588, CESifo, Munich, Germany).
FALK, A. and SZECH, N. (2013), "Morals and Markets", *Science*, **340**, 707–711.
FISCHBACHER, U. (2007), "z-Tree: Zurich Toolbox for Ready-made Economic Experiments", *Experimental Economics*, **10**, 171–178.
FISCHER, P., KRUEGER, J. I., GREITEMEYER, T. ET AL. (2011), "The Bystander-Effect: A Meta-analytic Review on Bystander Intervention in Dangerous and Non-dangerous Emergencies", *Psychological Bulletin*, **137**, 517–537.
FISMAN, R., KARIV, S. and MARKOVITS, D. (2007), "Individual Preferences for Giving", *American Economic Review*, **97**, 1858–1876.
FOOT, P. (1967). "The Problem of Abortion and the Doctrine of Double Effect", *Oxford Review*, **5**, 5–15.
GÄCHTER, S., NOSENZO, D., RENNER, E. ET AL. (2012), "Who Makes a Good Leader? Cooperativeness, Optimism and Leading-by-Example", *Economic Inquiry*, **50**, 953–967.
GÄCHTER, S., NOSENZO, D. and SEFTON, M. (2013), "Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?", *Journal of the European Economic Association*, **11**, 548–573.
GERT, B. (2012), "The Definition of Morality", in Zalta, E. N. (ed.), *Stanford Encyclopedia of Philosophy* (Fall 2012 ed.) (Stanford, CA: The Metaphysics Research Lab).
GINO, F., NORTON, M. I. and WEBER, R. A. (2016), "Motivated Bayesians: Feeling Moral While Acting Egoistically", *Journal of Economic Perspectives*, **30**, 189–212.
GLOVER, J. and SCOTT-TAGGART, M. (1975), "It Makes no Difference Whether or Not I Do It", *Proceedings of the Aristotelian Society, Supplementary Volumes,* **49**, 171–209.
GREENE, J. D., NYSTROM, L. E., ENGELL, A. D. ET AL. (2004), "The Neural Bases of Cognitive Conflict and Control in Moral Judgment", *Neuron,* **44**, 389–400.
GREINER, B. (2004), "An Online Recruitment System for Economic Experiments", in Kremer, K. and Macho, V. (eds) *Forschung und wissenschaftliches Rechnen*, Volume 63 of *GWDG-Bericht*. Göttingen, Germany: Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), 79–93.
HAMMAN, J. R., LOEWENSTEIN, G. and WEBER, R. A. (2010), "Self-interest through Delegation: An Additional Rationale for the Principal-Agent Relationship", *American Economic Review*, **100**, 1826–1846.
KANT, I. (1996), "Groundwork of the Metaphysics of Morals", in Gregor, M. J. (ed.), *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (Cambridge, UK: Cambridge University Press) 37–108.
LATANÉ, B. and DARLEY, J. M. (1968), "Group Inhibition of Bystander Intervention in Emergencies", *Journal of Personality and Social Psychology*, **10**, 215–221.
LIFTON, R. J. (1986), *The Nazi Doctors: Medical Killing and the Psychology of Genocide* (New York, NY: Basic Books).
MILGRAM, S. (2009), *Obedience to Authority: An Experimental View* (Reprint ed.) (New York, NY: Harper Perennial Modern Classics).
NORTON, M. I., MOCHON, D. and ARIELY, D. (2012), "The IKEA Effect: When Labor Leads to Love", *Journal of Consumer Psychology*, **22**, 453–460.

PECK, T. (2016), "If We Don't Sell Arms to Saudi Arabia, Someone Else Will, Says Boris Johnson", *The Independent October 26*, https://www.independent.co.uk/news/uk/politics/if-we-dont-sell-arms-to-saudi-arabia-someone-else-will-says-boris-johnson-a7382126.html (accessed 23 January 2020).
QUATTRONE, G. A. and TVERSKY, A. (1984), "Causal Versus Diagnostic Contingencies: On Self-deception and on the Voter's Illusion", *Journal of Personality and Social Psychology*, **46**, 237–248.
ROEMER, J. E. (2010), "Kantian Equilibrium", *Scandinavian Journal of Economics*, **112**, 1–24.
ROEMER, J. E. (2015), "Kantian Optimization: A Microfoundation for Cooperation", *Journal of Public Economics*, **127**, 45–57.
ROTHENHÄUSLER, D., SCHWEIZER, N. and SZECH, N. (2018), "Guilt in Voting and Public Good Games", *European Economic Review*, **101**, 664–681.
SERRA-GARCIA, M. and SZECH, N. (2018), "The (In)Elasticity of Moral Ignorance" (Working Paper Series in Economics 120, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany).
SOBEL, J. (2010), "Do Markets Make People Selfish?" (Mimeo).
SOLTES, E. (2016), *Why They Do It: Inside the Mind of the White-Collar Criminal* (New York, NY: PublicAffairs).
SPRANCA, M., MINSK, E. and BARON, J. (1991), "Omission and Commission in Judgment and Choice", *Journal of Experimental Social Psychology*, **27**, 76–105.
THOMSON, J. J. (1976), "Killing, Letting Die, and the Trolley Problem", *The Monist*, **59**, 204–217.