Looking Bad or Feeling Guilty?
The Effects of Intentions and Outcomes on Memory

Aurélien Klopfenstein*

September 13, 2026

[Link to latest version]

Abstract

The moral costs associated with selfish behavior are thought to explain in large part human prosociality, but recent experimental studies have shown that people can sometimes avoid the cost of remorse by selectively forgetting about an episode. This paper studies how two dimensions of moral costs, image concerns and concerns about harm, shape the selectivity of recall. In a preregistered two-session online experiment ($N = 680$), we use a modified dictator game involving risk to decorrelate the selfishness of actions from their outcomes for others. Our main finding is that harmful outcomes are especially memorable for participants who acted selfishly. The data rule out a number of purely cognitive explanations, but provide evidence for a guilt mechanism: the experience of guilt is sensitive to the actual harm caused to the third-party, and is associated with enhanced recall among selfish participants. We articulate these findings with the existing literature on selective forgetting based on image concerns.

*Paris School of Economics, École Normale Supérieure, PSL. I am very grateful to Nicolas Jacquemet and Maël Lebreton for their remarkable guidance and support. I also thank Alexander Cappelen, Evan Friedman, Charlotte Saucet, Erik Sørensen, Morten Støstad, Bertil Tungodden, Joël van der Weele and Roberto Weber for valuable discussions, as well as participants in the Paris Behavioral Economics Symposium, the Annual Conference of the French Experimental Economics Association, and the TIBER Symposium for helpful comments. The experiment was preregistered following the AsPredicted preregistration template; the preregistration is available at https://osf.io/8q5xb. The experimental procedures received approval from the Paris School of Economics Institutional Review Board (#2025-011). This work has been funded by a French government subsidy managed by the French National Research Agency (ANR) under the framework of the "Investments for the Future" program (reference ANR-17-EURE-001).

> "Such is the nature of that sentiment, which is properly called remorse; of all the sentiments which can enter the human breast the most dreadful. It is made up of shame from the sense of the impropriety of past conduct; of grief for the effects of it; of pity for those who suffer by it; and of the dread and terror of punishment from the consciousness of the justly provoked resentment of all rational creatures."
>
> - Adam Smith, *The Theory of Moral Sentiments* (1759/2022), p. 76.

1 Introduction

The experience of remembering having done something wrong or having harmed someone can be extremely unpleasant. This fact is well illustrated by the character of Raskolnikov in the novel *Crime and Punishment*, who, after committing a murder, endures psychological distress so intense that it eventually leads him to confess his crime (Dostoyevsky, 1866/2006).

It is thus no surprise that people would use whatever degrees of freedom they might have in order to avoid experiencing such memories. Indeed, in certain experimental settings, people have been found to remember their altruistic actions more accurately than their selfish ones (Saucet & Villeval, 2019), and to recall being more generous in the past than they actually were (Carlson et al., 2020). These findings contribute to a growing economic literature on "motivated forgetting" showing that people have a tendency to suppress or alter memories which put them in a bad light (for a review, see Amelio & Zimmermann, 2023). This tendency manifests itself not only in laboratory experiments, but also in field experiments and observational data, with practical implications in topics as diverse as managers' overconfidence (Huffman et al., 2022), rationalization of big life choices (Mueller, 2022), or health-relevant decisions (Sial et al., 2023).

Yet motivated forgetting must be limited in some ways, especially in moral contexts, or else moral motivations would lose most of their bite. In many economically-relevant social interactions, such as when deciding whether to help a family member, to shirk some responsibilities at work, or to lie when filing taxes, moral motivations may drive individual decisions to a similar or greater extent than classical incentives. In these interactions, the availability of motivated forgetting might determine when moral motivations really weigh on behavior, and when people will be able to simply brush them off.

When, then, should we expect remorse to be quickly forgotten, and when will it stick around? This paper adresses this question by testing experimentally the hypothesis that the content of remorse can make it easier or harder to suppress. As the opening quote suggests, remorse is not monolithic: it can include both "shame" about having done something bad intentionally, as well as concerns about the action's harmfulness for others:

the "grief for the effects of it" or "pity for those who suffer by it" (Smith, 1759/2022). This distinction appears in recent models of moral preferences in behavioral economics (Bénabou & Tirole, 2011; Bénabou et al., 2023), which differentiate between a basic aversion to others' being harmed, and a signalling or self-image motive, which captures the shame that people experience when their actions reflect poorly on their intentions and their moral character, no matter the final consequences.

Why would image concerns and concerns about harm interact differently with memory? Theoretical accounts of motivated forgetting attribute it mainly to image concerns (Bénabou & Tirole, 2002, 2011), in line with observations about shame and avoidance in psychology (Sznycer, 2019). This leaves room for a distinct effect of concerns about harm on memory, which we investigate empirically. Importantly, existing studies of the memory of social interactions could not isolate such an effect, because they are based on dictator games in which acting selfishly and harming a third-party are perfectly correlated (Saucet & Villeval, 2019; Carlson et al., 2020)[1].

In order to decorrelate these two dimensions of moral preferences, we use a variant of the dictator game which introduces random chance between the dictator's action and the recipient's outcome, as described in Figure 1. While we introduce this game specifically to isolate a potential effect of concerns for outcomes on memory, the general idea that people can sometimes gain from exposing others to a risk has some relevance to real-world situations such as that of a financial advisor, or a quality-control worker. In the experiment, participants play the game in an initial session, and come back forty-eight hours later for a follow-up session in which they have to recall both what their action was, and what the recipient's outcome was, thus summarizing their beliefs about what happened during the initial session.

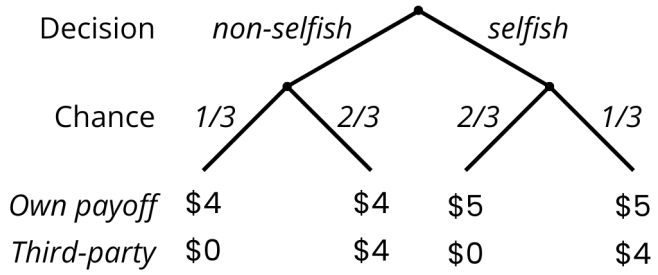


Figure 1: Dictator game variant involving risk.

A first outcome from the experiment is to provide a conceptual replication of existing studies of motivated memory based on image concerns in dictator games (Saucet & Villeval, 2019; Carlson et al., 2020). Our main test of motivated memory is to compare the dictators'

[1]Also note that participants in these studies took part in a series of social interactions, of which only one was incentivized, potentially making the sense of a specific harm caused to the third-party more diffuse. In Saucet and Villeval (2019), participants played twelve binary dictator games. In Carlson et al. (2020), they played five continuous ones.

recall accuracy depending on what their action was. Because we hypothesize that harm to the recipient may have an additional effect on memory, the cleanest comparison is between dictators whose recipient was not harmed.

Surprisingly, we find quite mixed evidence for motivated forgetting on the dictators' part. In particular, dictators who acted selfishly are more, not less, likely to correctly recall their action, no matter the recipient's outcome, although this tendency is only directional. On memory of the outcome, we do find that when the recipient was not harmed, dictators who acted selfishly are 16 percentage points less likely to correctly recall the outcome than dictators who acted altruistically. However, the significance of the effect is not very robust across regression specifications, especially when controlling for differences in a measure of episodic memory ability.

These mixed findings contrast with the clear results of Saucet and Villeval (2019) and Carlson et al. (2020), which provided the first evidence of motivated forgetting in dictator games. Our experiment is not an exact replication of either of these studies, which are based on different social interactions, and elicit memories differently than we do. However, these differences open the possibility that motivated forgetting is sensitive to certain experimental parameters, and is thus less prevalent than previously thought.

The main contribution of this paper is to study the effect of concerns about harm on memory. On memory of the outcome, the experimental data provide strong evidence that being responsible for causing harm leads to a much higher rate of recall of the outcome. We find that dictators who acted selfishly have very accurate memories of the outcome when the recipients were harmed, and rather inaccurate ones when they were not. On memory of the action, we again do not find any significant differences, and the estimated effects are very small. Importantly, it indicates that dictators who acted selfishly did not remember their action differently when it had harmful consequences.

The empirical pattern of the dictators' memories allows us to rule out a number of purely cognitive mechanisms which could in principle explain why harmful outcomes are so well remembered by selfish dictators. One possibility would be that the recipient losing the endowment is a more negative or extreme outcome, making it memorable (Baumeister et al., 2001). Yet we find that dictators who acted altruistically are not more likely to recall harmful outcomes, but slightly less. Another explanation might be that dictators have a tendency to recall whichever outcome was more likely given their action, as a result of a probabilistic inference. However, the joint distribution of memories of the action and of the outcome does not fit such a pattern, and instead reflects two distinct effects: among non-selfish dictators, correct recall of the action is correlated with correct recall of the outcome, no matter what it was; among selfish dictators, harmful outcomes are always highly memorable, no matter what the recalled action is.

Having excluded a number of non-moral interpretations of the findings, we turn our attention to mechanisms related to moral preferences. Specifically, we hypothesize that a

prominent one should be that dictators experience the emotion of guilt after causing harm to the recipients. Psychological accounts of guilt have proposed that experiencing guilt serves various adaptive functions, such as motivating individuals to engage in reparatory actions (Baumeister et al., 1994; Tangney & Dearing, 2002), or inducing long-term learning (Tooby & Cosmides, 2008). Given these functions, we argue that the feeling of guilt should be sensitive to the actual harm caused to others, and associated with enhanced memory.

In the experiment, after seeing the outcome of the game, dictators report the extent to which their participation in the session made them feel several emotions, including guilt, which allows us to test these hypotheses. We find that guilt is indeed sensitive to outcomes: dictators report feeling significantly more guilty when recipients were harmed, compared to when they were not. Perhaps surprisingly, this is the case for all dictators, no matter their action, a finding which we relate to the known association between guilt and empathy in psychological questionnaires (Tangney & Dearing, 2002). Finally, in the absence of any harm, having acted selfishly increases guilt only directionally. Thus, we find strong support for the idea that guilt is associated with moral concerns about harm.

We then test whether the emotion of guilt is responsible for the memorability of the harm caused to the recipient. In this regard, we find that reported guilt robustly predicts enhanced recall of the outcome across a variety of specifications, but only among dictators who acted selfishly, and are thus responsible for the outcome. An extensive literature in psychology argues that emotions, and in particular negative emotions, tend to be memorable (see e.g. Brown and Kulik, 1977; for reviews, see Kensinger, 2007; McGaugh, 2013). However, in our study, the link between responsible guilt and recall appears to be very specific: none of the other reported emotions are associated with memory, and reported guilt does not predict recall among dictators who acted altruistically and are thus not responsible for the outcome. These results highlight the importance of studying the functional relationships of memory with specific emotions, and not just with emotions in general (Levine & Pizarro, 2004).

Can causing small harm to someone alter an individual's behavior days later? Our design allows us to study the consequences of having caused harm on downstream behavior. In the follow-up session, participants complete two behavioral tasks, one meant to capture the role of guilt in reparation, and the other its role in learning. Briefly, we do not find any effect of recipients' outcomes on the dictators' downstream behavior, with very small and non-significant effects in both tasks. Of course, these null results do not mean that moral concerns about harm have no behavioral effects. In fact, we do stumble upon an unexpected one: dictators who acted selfishly are slightly but significantly less likely to take part in the follow-up session when they caused harm to the recipient. This differential attrition does not alter the previous conclusions, which are robust to conservative assumptions about the missing dictators. However, it implies that some dictators who acted selfishly to increase their bonus reward by 25% end up foregoing it entirely to avoid returning to the

"scene of the crime".

Overall, this paper contributes to delineating the limitations of motivated forgetting in the moral domain. It complements findings about domain-general limitations, such as the effect of recall incentives (Zimmermann, 2020; Gödker et al., 2025), or of participants being aware that their memory will be tested in the future (Zimmermann, 2020). In moral contexts, a few studies already highlight some limitations of motivated forgetting. For instance, in an experiment on dishonesty, Galeotti et al. (2020) do not find motivated forgetting in their baseline condition, but do find some when it provides a justification for further selfish behavior. In psychology, Stanley et al. (2018) replicate a study in which participants have to recall the details of a story which can involve or not a moral transgression (Kouchaki & Gino, 2016). Contrary to the results of the initial study, they do not find any effect of this variation on recall accuracy, and argue that memories of moral transgressions may become particularly accurate and vivid if they are associated with negative emotions. We concur, and offer experimental evidence that moral concerns about being responsible for harm, and guilt in particular, are what makes some moral transgressions especially memorable.

The remainder of the paper is structured as follows. In the next section, we describe the experimental design, before detailing our preregistered hypotheses within a simple framework of moral preferences in Section 3. Section 4 presents the results of the experiment on the dictators' side, and Section 5 provides several robustness checks. Finally, Section 6 concludes.

2 Experimental Design

The experiment takes place over two sessions with a 48-hour delay in between. In the initial session, participants have a one-shot social interaction in the form of a dictator-game variant, the memories of which we elicit in the follow-up session. In what follows, we describe the experimental design from the dictators' point of view, given that they are the main focus of this paper. Appendix A.5.1 describes concisely how instructions differ on the recipient side, while Appendix B contains the full instructions for dictators and recipients. The experiment was preregistered following the AsPredicted template. The preregistration is available at https://osf.io/8q5xb.

2.1 Initial session

The initial session opens with a psychometric test aiming to measure the participants' episodic memory ability[2]. After the test, participants receive a $4 endowment as a reward

[2]The task is an adaptation of the Picture Sequence Memory Test from the NIH Toolbox Cognition Battery, in which participants have to order a series of fifteen pictures in the sequence in which they previously appeared (Dikmen et al., 2014). Appendix A.3 provides additional details on our measures of

for completing this part. Then, they read the instructions for the dictator-game variant, before learning their roles. The game is as follows: participants play in pairs of one dictator (afterwards, "she") and one recipient ("he"), with fixed roles, which have neutral labels in the experiment ("Player 1/2"). The recipient is at risk of losing his \$4 endowment, depending on a draw from an urn containing blue and red balls. Drawing a red ball is the "bad" outcome for the recipient, which leads to the endowment being lost; drawing a blue ball is the "good" outcome, meaning that the recipient keeps the endowment.

The dictator has to make a decision which determines the share of red balls in the urn, corresponding to the probability p that the recipient loses the endowment. There are two possible options: $p = 1/3$ or $p = 2/3$. If the dictator chooses the higher probability of loss $p = 2/3$, she earns an unconditional \$1 bonus on top of her endowment. To make sure that dictators encode the information of the exact number of red balls corresponding to their choice, they have to select their preferred option by typing the corresponding number of red balls. We randomize at the pair level the meaning of colors, the order of options, as well as the total number of balls in the urn between three possibilities (21, 24 or 27 balls in total). The dictator's action is made common knowledge, as is the outcome of the draw and the resulting payoffs, which both players see for a fixed duration of around twenty seconds.

The design of the game meets several constraints. First, we aim to make the social interaction as consequential as possible: participants know that they play the game only once, with relatively high stakes compared to the short duration of the experiment. Second, we make the choice of a higher probability of loss unambiguously selfish: choosing the higher probability of loss decreases the expected sum of payoffs, always increases inequality, and cannot be justified by reciprocity concerns given that players interact only once and with a determined role. Third, the selfish gain of \$1 is attractive enough that a large share of dictators will still choose it. Finally, the symmetry of the lotteries rules out a number of purely cognitive factors influencing memory by design.

The initial session closes with a questionnaire recording the participants' age, gender, and self-reported risk attitude (Dohmen et al., 2011), followed by questions asking them to rate on a zero-to-ten scale to what extent their participation in the experiment made them feel four different emotions: happiness, guilt, pride and annoyance ("I have felt happiness/guilt/pride/annoyance": 0 = "Not at all", 10 = "Very much")[3].

2.2 Follow-up session

The follow-up session starts with the elicitation of the participants' memories of the dictator-game variant. They read a verbal description of the game reminding them of its

episodic memory ability.

[3]We restrict ourselves to only four emotions, compared to twenty in a more standard scale like PANAS (Watson et al., 1988), in order to avoid losing the attention of participants.

main features, and in particular that they had to choose between two symmetrical urns, that they had an incentive of $1 to choose the one with a higher proportion of red balls, and that the recipient would lose $4 if a red ball was drawn from the urn. The exact description is available with the instructions in Appendix B.2.

We then elicit the participants' memories of two specific details of the game: one corresponds to the memory of the action, the other to the memory of the outcome. Assuming that participants read the information provided, the two measures are sufficient to reconstruct the complete memory of the social interaction. The order of the two questions is randomized, and one of them is incentivized with a $0.5 bonus reward.

Regarding memory of the action, we ask participants to recall the exact number of red balls in the urn that they chose, incentivizing exact recalls. Because we remind participants of the total number of balls in the urn, and that the urns were symmetrical, the number of red balls in the chosen urn maps onto the probability of loss for the recipient, and thus to how selfish the dictator's action was. We ask participants to freely recall the exact number of red balls in the urn, instead of picking them out of existing options, in order to limit the participants' ability to recover what their choice was simply by making their decision again. Finally, we ask participants to rate their confidence in their answer on a seven-point scale.

Regarding memory of the outcome, we ask participants to recall the color of the ball that was drawn. Because we remind participants of the consequences of each color, it maps onto the recipient's payoff. After having selected which ball they recall was drawn, participants also report their subjective probability of being correct, from 50% to 100%. We incentivize this reported probability with a probabilistic Becker-DeGroot-Marschak mechanism which is incentive-compatible even under risk aversion (Karni, 2009)[4].

Finally, after having recorded their memories of the game, participants complete another memory task in which they have to correctly recall the picture stimuli used in the task measuring memory ability in the initial session. We use this measure to test whether the social interaction may influence the memory of an unrelated but temporally close task, and as a secondary measure of memory ability (see Appendix A.3 for details).

In the follow-up session, participants also complete in a randomized order two behavioral tasks which aim to assess whether the dictator-game variant from the initial session, and the associated memories, have downstream influences on their behavior. One of the tasks is a third-party allocation task, borrowed from Gurdal et al. (2013), in which participants state how they would like to allocate 15$ between their partner from the dictator game, and another participant who had the same role as their partner. Participants are informed that the researcher will implement through bonus payments the allocation decision of

[4]Participants are told that it is in their best interest to report the real chances from their point of view, and can click a button to get an intuitive but detailed explanation of the procedure (adapted from Hao and Houser, 2012).

one participant in the entire experiment. This task aims to capture forms of reparation towards the recipient on the dictator side, or forms of punishment of the dictator on the recipient side. As in Gurdal et al. (2013), we use allocation between the partner and a third-party to avoid potential wealth effects which could confound the comparisons in an allocation between self and partner. Importantly, the symmetry of our design for recipients means that we also offer a conceptual replication and extension of Gurdal et al. (2013); Appendix A.5 presents the corresponding results.

The other behavioral task is a multiple-price-list version of the initial dictator game, which aims to capture a potential learning effect. The game involves again two urns with different expected values for a recipient ($4 or $8). All participants state which urn they would choose depending on the value of the selfish gain associated with choosing the low expected-value urn (from $8 to $13). Again, participants are informed that the researcher will implement one of those actions, for exactly one participant in the entire experiment, with a randomly selected recipient.

Finally, in order to measure the test-retest reliability of emotion ratings, we ask participants to report again on zero-to-ten scales the extent to which their participation in the first session made them feel happiness, guilt, pride and annoyance. Participants also report whether they have thought again about the initial session since they completed it, how much time they spent on Prolific since the initial session, and whether they enjoyed taking part in the experiment.

2.3 Procedures

The data was collected online in January 2026, with participants recruited on Prolific, for a total sample of 680 US participants (351 women). Around the time of data collection, Prolific was estimated to have a low rate of AI submissions (Celebi et al., 2026). The data was collected in two waves one day apart. For each wave, the recruitment for the initial session started at the same time of day and took place over an 8-hour time window. The recruitment for the follow-up session opened at a fixed time for all participants, 48 hours after the beginning of the recruitment for the initial session.

In order to minimize attrition between the two sessions, we took several precautions in the recruitment ad: we informed participants of the exact time at which the follow-up session would be available, asked them not to take part in the initial session if they did not think they would be able to complete the follow-up session, and made the bonus from the initial session conditional on completion of the follow-up. Participants were thus informed that there would be a follow-up session, but we did not specify what would happen in it. In order to check that participants did not prepare for a follow-up which would involve remembering the initial session, we ask them to report at the end of the follow-up whether they did anything to help them recall the initial session.

When the follow-up opened, we sent messages to the participants reminding them to complete it. After 8 hours, we sent another reminder to participants who still had not completed it. Because the experiment was conducted in the evening, participants who completed the experiment after the initial eight hours were likely to have an additional night of sleep, which might affect their memory. The main text reports the results for participants who completed the follow-up within 24 hours, and Appendix A.2 reports the results for participants who completed within 8 hours.

Of the 680 participants in the initial session (340 dictators), 618 or 90.9% completed the follow-up session within 24 hours of its opening (310 dictators). Of those, 555 or 89.8% complete it within the 8-hour window (274 dictators). Due to a technical error, 28 participants were disconnected from the follow-up session at some point after having reported their memories of the dictator game variant, leading to missing data on some variables of the follow-up session.

The target sample size of 300 dictators completing the follow-up session was determined from a power analysis, so as to detect with 90% power an effect of outcome on the dictators' outcome recall which would equal two-thirds of the effect observed in a pilot experiment.

The median duration for the initial session was 11 minutes, and participants received a fixed payment of $1.71-1.78 for their participation (the fixed payment was slightly increased for the second wave). The median duration for the follow-up session was 9 minutes, and participants received a fixed payment of $1.68 for their participation. The average bonus payment was $3.53. Participants who were disconnected due to a technical error were paid the fixed and bonus payments.

The experiment received the approval of the Paris School of Economics Institutional Review Board. Participants provided informed consent at the beginning of each experimental session.

3 Conceptual framework and hypotheses

3.1 Moral preferences

We articulate our hypotheses within a basic framework of moral preferences which integrates both image concerns, and concerns about harm. The model is not novel; it is closest to that of Bénabou et al. (2023) under direct elicitation and a continuum of types, and generally analogous to other models of signalling and moral concerns (see e.g. Bénabou and Tirole, 2006, 2011). We introduce it only as a framework for our investigation of memory.

In the experiment, a dictator i has to make a action $a_i \in \{0, 1\}$, with consequences on her own payoff $\pi_i(a_i)$, and on the recipient's payoff $\pi_j(a_i, \omega)$, which also depends on the state of nature $\omega \in \Omega$. $a_i = 0$ is the selfish action, meaning that $\pi_i(0) > \pi_i(1)$; we assume

that the non-selfish action stochastically dominates the selfish one from the point of view of the recipient (as is the case in the experiment): $\forall s > 0, \forall \omega \in \Omega, \max(0, s - \pi_j(0, \omega)) < \max(0, s - \pi_j(1, \omega))$. We assume that the dictator's utility U_i takes the following form:

$$U_i(a_i, \omega) = \pi_i(a_i) - \theta_i \max(0, s - \pi_j(a_i, \omega)) + \mu \mathbb{E}(\theta_i \mid a_i) \tag{1}$$

Thus, in addition to valuing her own payoff, the dictator's moral concerns are twofold. First, she experiences "grief for the effects" of her action, and "pity for those who suffer by it", in the form of an empathy term $-\theta_i \max(0, s - \pi_j(a_i, \omega))$, where θ_i is the dictator's empathy, and s is some fairness standard corresponding to what the recipient deserves as minimal payoff after the social interaction. We assume that dictators vary in their empathy θ_i, which follows a uniform distribution over $[0, \theta_{\max}]$.

Broadly, this term captures the dictator's concern for the consequences of her actions on the recipient, and specifically the harm compared to the fairness standard. In the context of the present experiment, the presence of a shared fairness standard has no influence on the equilibrium, as long as $s > 0$; however, it does shape the interpretation of the model, from which we derive our behavioral hypotheses. Specifically, it means that the dictator cares about not harming the recipient, but that she does not feel anything in particular when the recipient is not harmed. The model is similar to models of advantageous inequality aversion or guilt aversion, in that the agent does not incur costs when the payoffs match or exceed the fairness standard (Fehr & Fischbacher, 2002; Battigalli & Dufwenberg, 2007). As a simplifying assumption, we assume the fairness standard s to be equal to the participants' starting endowment in the game. The fact that this term is dependent on the realized outcome does not change the dictators' actions, because they maximize expected utility; but it has implications for the predictions on *ex post* behavioral effects.

The dictator also cares about the "shame from the sense of the impropriety of past conduct". Leading accounts of shame in evolutionary psychology have proposed that it is a self-conscious emotion (Sznycer, 2019; Tangney et al., 2007), which closely tracks the threat of devaluation by others (Sznycer et al., 2016). We thus model shame as a signalling motivation, corresponding to the fact that the dictator generally wants to signal good intentions, and feels shame when her actions signal bad intentions. In the model, the dictator wants to act in a way that signals having a high θ_i, weighted by a context-specific parameter μ. This interpretation of the signalling term as corresponding to the emotion of shame sits half-way between the two classical interpretations of such signalling terms, as either social signalling or self-signalling (see e.g. the discussion in Bénabou & Tirole, 2006). Shame is an internal indicator, but it reflects what the judgments of others would be, would they observe the individual's behavior, and helps the individual monitor his reputation. In particular, even in the absence of any observer, an individual might feel ashamed of an action, which may in fact motivate her to further conceal her action to the

eyes of others (Sznycer, 2019). However, our experiment does not allow to differentiate these interpretations, so all three are admissible: note that both the recipient, and the experimenter know the dictator's action. Whichever interpretation is preferred, the focus of this paper is on the fact that such a signalling term captures an intentional dimension of moral preferences, which does not depend on the final payoff of the recipient, or on the realization of ω.

In the case where dictators share the same fairness standard, the model naturally leads to a cutoff $\bar{\theta}$, above which dictators choose the non-selfish action $a_i = 1$, while they choose the selfish action $a_i = 0$ below it. Thus, dictators who choose $a_i = 1$ experience an image of $\mathbb{E}(\theta_i \mid \theta_i > \bar{\theta})$ and those who choose $a_i = 0$ experience an image of $\mathbb{E}(\theta_i \mid \theta_i < \bar{\theta}) < \mathbb{E}(\theta_i \mid \theta_i > \bar{\theta})$. The details of the resolution are available in the Appendix of Bénabou et al. (2023).

3.2 Hypotheses

We now present our preregistered hypotheses, starting with those regarding memory. Recall that we elicit memories of two details relevant to the social interaction in the dictator-game variant: the dictator's action, and the recipient's outcome. We do not treat these two details differently, and make the same conjectures on the recall rates of each.

The first hypothesis bears on the effect of image concerns on the recall of the social interaction. Because individuals want to appear well-intentioned towards others, they may be motivated to remember accurately episodes in which they were, or complementarily, to forget episodes in which they were not. To test this effect in the experiment, we compare the memories of dictators who made different actions, but obtained the same outcome for the recipient. Given that we also hypothesize that bad outcomes have their own effect on memory, the cleanest comparison is among dictators whose recipients did not lose their endowment. One potential concern with this comparison is that selfish and non-selfish dictators may differ in their episodic memory abilities; we mitigate it by controlling for our measure of memory ability.

H1 (motivated forgetting) – When the recipient did not lose the endowment, dictators who acted selfishly are less likely to correctly recall details of the social interaction than dictators who made the non-selfish choice.

Hypothesis 1 is in harmony with the existing findings that people sometimes tend to forget the details of selfish episodes (Saucet & Villeval, 2019), especially when their actions violated their fairness standards (Carlson et al., 2020). In contrast, the next hypotheses rely on this specificity of our experimental setting as they bear on effects of the recipient's outcome on the dictator's memory, experience and subsequent behavior. The dictator's concerns over not harming the recipient lead her to experience a cost when she does. Let

us consider two hypotheses regarding the effect of such cost on the dictator's memory of the episode. First, the emotion of empathy, materializing this cost for the dictator, might lead her to recall more accurately the episode when the recipient was harmed. An extensive literature documents that episodes associated with emotions are more vividly remembered (McGaugh, 2013), leading to the following hypothesis:

H2 (empathic recall) – Dictators are more likely to correctly recall details of the social interaction when the recipient lost his endowment, compared to when the recipient did not lose the endowment.

However, we also consider the alternative possibility that empathy alone is not sufficient to enhance recall. Specifically, we preregistered the alternative hypothesis that it is only when the dictator acted selfishly, and is therefore responsible for harming the recipient, that bad outcomes lead to more accurate recall of the episode. To make this prediction, we rely on functional accounts of the emotion of guilt, which propose that the experience of guilt after wrongdoing depends on the outcome and that it motivates reparatory behavior, and learning (Baumeister et al., 1994; Tooby & Cosmides, 2008; Fitouchi et al., 2024). An individual who did the right thing would not need to make reparation for accidental harm, nor learn from the outcome. This leads us to the following hypothesis:

H2' (guilty recall) – Dictators who acted selfishly, but not those who did not, are more likely to correctly recall details of the social interaction when the recipient lost his endowment, compared to when he did not.

The next hypotheses aim to shed light on one potential mechanism behind the sensitivity to outcomes, which is that the experience of guilt is itself outcome-sensitive. Again, we preregistered two alternative hypotheses, one which concerns all dictators, the other which is restricted to dictators who acted selfishly:

H3 (outcome-sensitive guilt) – Dictators experience and report more guilt when the recipient lost his endowment, compared to when he did not.

H3' (responsible guilt) – Dictators who acted selfishly, but not those who did *not*, experience and report more guilt when the recipient lost his endowment, compared to when he did not.

If functional accounts of guilt are to be trusted, then the differences in experienced guilt and in memory depending on the recipient's outcome should translate into differences in subsequent behavior. Based on the behavioral measures available in the second session, we make the following hypothesis:

H4 (behavioral consequences) – After the recipient has lost the endowment, dictators engage in more reparative behavior and act more prosocially, compared to when he has not.

Appendix A.1 details the correspondence between these hypotheses and the preregistration hypotheses, as well as the results of the preregistered statistical analyses. The preregistration includes one additional hypothesis about motivated memory among recipients, which we do not report on in the main text for concision. The complete results on the recipient side of the experiment are reported in Appendix A.5.

4 Results

In this section, we report the results of the experiment on the dictators' side. We first report on the dictators' behavior in the initial session. Then, we study their memories of the action and of the outcome, as well as the role of guilt in memory. Finally, we give the results on the downstream behavioral measures. For ease of presentation, we present both preregistered and exploratory results together; Appendix A.1 provides the exact preregistered analyses and lists deviations from the preregistration.

4.1 Behavior in the initial session

The dictator game variant aims to create a social dilemma for dictators: will they increase by 1/3 the probability that the recipient loses his \$4 endowment, in order to secure a selfish \$1 bonus for themselves? In the data, 49.4% of dictators choose the selfish option. Consistent with dictators weighing the implications of their action, the median decision time is 25.3 seconds; decision times are similar for dictators who end up choosing the selfish or the non-selfish option (median for selfish: 27.2"; median for non-selfish: 24.5"; exact Kolmogorov-Smirnov test of differences in the distribution of decision times: $D = 0.131, p = 0.096$).

Table 1 shows various demographic and behavioral characteristics of selfish and non-selfish dictators. We do not observe systematic gender or age differences between these groups. Dictators who act selfishly are however more risk-seeking on average, consistent with the fact that choosing the selfish option is a form of social risk-taking. An important observation, given the topic of our study, is that dictators who act selfishly are on average 0.21 standard deviations lower in our measure of episodic memory ability, although the trend is non-signifiant (unequal variances t-test: $t(328.52) = -1.96, p = 0.051$). Accordingly, we present our results comparing dictators who acted differently with and without this measure of episodic memory ability as control. For an extended discussion of our measures of episodic memory ability, and additional checks of the results' robustness to including

them as control, see Appendix A.3.

Table 1: Characteristics of selfish and non-selfish dictators

	Non-selfish (N=171)		Selfish (N=162)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Women	0.53	0.50	0.50	0.50	-0.03	0.05
Age	37.90	9.59	38.38	9.60	0.48	1.05
Risk attitude	4.73	2.46	5.64	2.61	0.90**	0.28
Memory ability	0.05	0.97	-0.16	1.00	-0.21	0.11
Decision time (s)	29.61	20.93	34.60	31.17	4.98	2.93

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Summary statistics of selfish and non-selfish dictators. Memory ability is the normalized episodic memory performance measured in the initial session; we exclude dictators who reported having cheated on the memory task.

4.2 Memories of the action

We now report the results for our hypotheses on the dictators' recall of their actions, which they report as the number of red balls in their chosen option. Figure 2a sums up the result for exact recall of the number of red balls. We find that dictators have similar rates of exact recall of their action, no matter their action, or its outcome for the recipient. Directionally, selfish dictators are 8.9 percentage points more likely than non-selfish dictators to correctly recall their actions, contrary to our prediction in H1, although the difference is not significant (χ^2-test of difference of proportion between selfish and non-selfish dictators: $\chi^2(1) = 2.53, p = 0.112$). Table 2 shows the results in regression form.

For robustness, we also give the results for recall of the action's selfishness, meaning that the participant correctly recalls having chosen an urn with a majority or a minority or red balls [5]. The description on the elicitation screen makes it clear that the dictator earned the selfish bonus when the urn contained a majority of red balls. Recall of the action's selfishness is possibly more meaningful for participants than exact recall, given that the exact number of red balls may not have been the decisive factor in the dictator's decision. For instance, a dictator who acted selfishly may have chosen the option offering the selfish bonus without paying any attention to the corresponding urn composition. However, this advantage comes with the previously mentioned disadvantage that some dictators may know through introspection that they would always choose the selfish or the non-selfish option, meaning that the measure may be confounded by preference stability.

[5] We consider the three participants who report recalling an urn containing exactly 50% of red balls as incorrect about their action's selfishness.

The results for recall of the action's selfishness are almost identical to the results for exact recall. Results are also identical when defining recall as reporting a number within a distance of one or two from the true number of red balls.

Table 2: Recall of the action among dictators

	Exact recall		Selfishness recall	
	(1)	(2)	(3)	(4)
Constant	0.549***	0.546***	0.676***	0.674***
	(0.049)	(0.049)	(0.046)	(0.046)
Selfish choice (S)	0.095	0.101	0.057	0.061
	(0.087)	(0.090)	(0.081)	(0.084)
Bad outcome	-0.032	-0.031	-0.021	-0.021
	(0.082)	(0.082)	(0.078)	(0.077)
Bad outcome × S	0.006	0.016	0.040	0.044
	(0.119)	(0.120)	(0.110)	(0.112)
Memory ability		0.035		0.026
		(0.029)		(0.027)
Adjusted R^2	-0.001	0.001	-0.002	-0.002
Observations	310	306	310	306

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of exact recall of the action and correct recall of selfishness among dictators, with heteroskedasticity-consistent standard errors. Exact recall of selfishness is defined as correctly recalling the exact number of red balls in the urn. Correct recall of selfishness is defined as correctly recalling what color was the majority of balls in the urn. The sample includes dictators who completed the follow-up session within the 24-hour window. Columns (2) and (4) excludes those who reported having cheated on the memory task.

Result 1a – Contrary to H1, we do not find significant differences in recall of the action between selfish and non-selfish dictators, whether the recipient lost his endowment or not. Directionally, dictators who acted selfishly are more likely to correctly recall their action.

Result 2a – Contrary to H2 and H2', we do not find significant differences in recall of the action among dictators depending on the outcome for the recipient

Overall, our measures of recall of the action do not pick up any variability in the dictators' memory. Especially surprising is the fact that even episodic memory ability does not significantly predict recall of the action. Given that the rate of exact recall of the number of red balls is still very high (58.1% overall), it is possible that differences would emerge after a longer time had elapsed. Another possible explanation is that recalling

the exact number of red balls mainly depends on having encoded the information not as a number, but as a proportion, either 1/3 or 2/3. If this is the case, then differences in numeracy skills may in large part determine recall, which could explain why other plausible predictors do not play any role.

4.3 Memories of the outcome

While the number of red balls in the urn should influence the evaluation of how bad the dictator's action is, the color of the ball that is drawn is what matters in the end for the recipient's outcome. We now turn our attention to memories of this detail. Figure 2b shows the rate of correct recall of the outcome depending on the dictator's action and the recipient's outcome, and Table 3 the corresponding regressions.

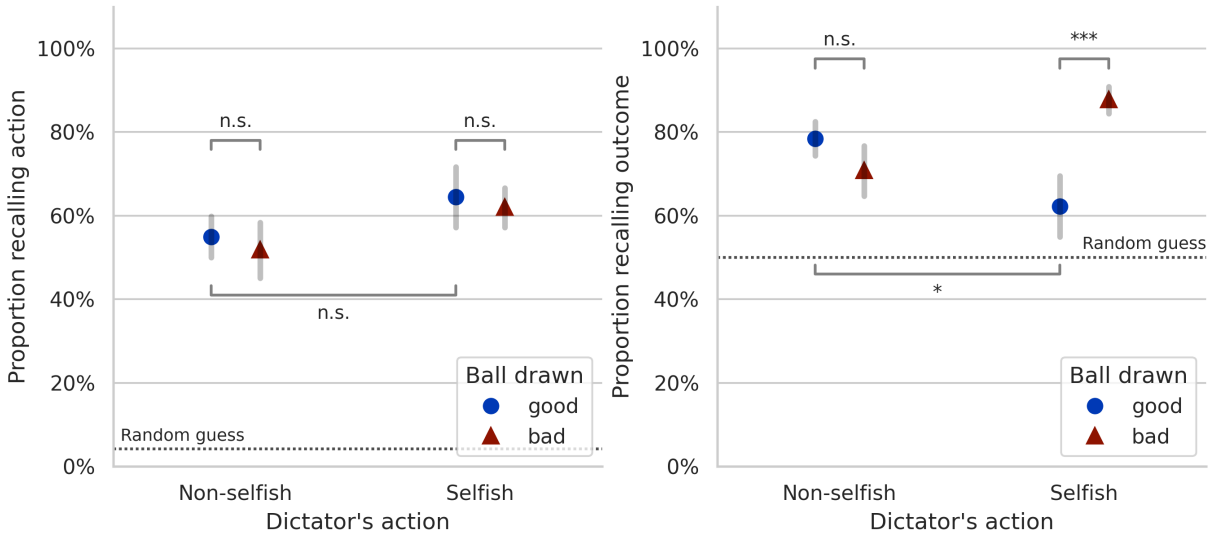
In support of the motivated forgetting hypothesis (H1), we find that when the recipient did not lose his endowment, dictators who acted selfishly are 16.2 percentage points less likely to correctly recall the outcome (χ^2 test of proportions: $\chi^2(1) = 4.21, p = 0.040$). However, this tendency should be interpreted with caution, as some statistical tests of the hypothesis, including the preregistered one, or others which include memory ability as a control, are not significant at the 5% level (see Appendix A.1). In addition, given that we assume that motivated forgetting is driven by intentional concerns such as self-image or shame, it may come as a surprise that we should observe motivated forgetting on outcomes, but the opposite direction on actions, which are more directly related to intentions.

Result 1b – When the recipient did not lose his endowment, dictators who acted selfishly tend to correctly recall the outcome less often than non-selfish dictators, providing some support for H1.

We now turn our attention to the effect of outcomes. We find that among dictators who acted altruistically, recall of the recipient's outcome is similar, no matter what it was (χ^2 test of proportion: $\chi^2(1) = 1.20, p = 0.273$; difference of 7.7 p.p.). However, dictators who acted selfishly are much more likely to correctly recall the recipient's outcome when it was bad, than when it was good (χ^2 test of proportion: $\chi^2(1) = 12.70, p < 0.001$; difference of 25.4 p.p.).

Result 2b – In support of H2', but not H2, dictators who acted selfishly are much more likely to correctly recall the outcome when the recipient lost his endowment, whereas dictators who acted altruistically recall both outcomes at similar rates.

A number of alternative explanations of Result 2b may come to mind. Notice, however, that a valid explanation needs to explain both why dictators who acted selfishly recall bad outcomes very well, and why this is not particularly the case for dictators who acted



(a) Recall of the number of red balls

(b) Recall of the ball's color

Figure 2: (a) Rate of exact recall of the number of red balls in the urn among dictators, depending on the dictator's choice and outcome. The dotted line represents the chance level with a uniform prior($\approx 4\%$). (b) Rate of correct recall of the recipient's outcome among dictators, depending on the dictator's choice and the outcome. The dotted line represents the chance level (50%).
Bars represent non-robust standard errors. Stars correspond to χ^2 tests of proportions: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

altruistically. For instance, the recipient's payoff of \$0, in a game where the possible payoffs are \$0, \$4 and \$5, could be considered extreme, and there is some existing evidence that extreme lottery outcomes are more memorable (Madan et al., 2014). However, such an explanation also predicts that dictators who acted altruistically should show the same pattern, so it fails to explain our results. The same reasoning applies to the argument that bad outcomes are negative events, and that negative events leave a stronger mark in memory (Baumeister et al., 2001). Arguably, if a bad outcome is a negative event for selfish dictators, it should be even more so for non-selfish dictators, who presumably care more about the recipient's payoff, and made a personal sacrifice which eventually proved useless.

One important explanation to consider is the possibility that dictators simply tend to recall what the most common outcome was, given their action. Dictators who acted selfishly are more likely to recall that the outcome was bad, while dictators who acted altruistically are more likely to recall that it was good. Such a pattern could appear simply because dictators tend to give coherent answers. For instance, dictators who recall what their action was, but not its outcome, may answer the question about the outcome with the most likely outcome given their action. More generally, participants could perform some kind of two-way probabilistic inference between their memories of the action, of the outcome, and their confidence in each, and answer the survey based on the resulting

Table 3: Correct recall of the outcome among dictators

	(1)	(2)	(3)	(4)
Constant	0.784***	0.779***	0.574***	0.576***
	(0.041)	(0.041)	(0.073)	(0.072)
Selfish choice (S)	-0.162	-0.149	0.016	0.015
	(0.083)	(0.085)	(0.117)	(0.119)
Bad outcome	-0.077	-0.076	-0.071	-0.070
	(0.072)	(0.071)	(0.069)	(0.068)
Bad outcome × S	0.331**	0.335**	0.324**	0.328**
	(0.107)	(0.108)	(0.105)	(0.105)
Correct about selfishness			0.311***	0.301***
			(0.076)	(0.076)
Correct about selfishness × S			-0.267*	-0.250*
			(0.107)	(0.108)
Memory ability		0.057*		0.049*
		(0.025)		(0.025)
Adjusted R^2	0.035	0.054	0.095	0.111
Observations	310	306	310	306

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of correct recall of the outcome among dictators, with heteroskedasticity-consistent standard errors. The sample includes dictators who completed the follow-up session within the 24-hour window. Columns (2) and (4) exclude those who reported having cheated on the memory task.

computation.

Under this explanation, we should expect the inverse pattern among dictators who incorrectly recall their action's selfishness. For instance, a dictator who acted selfishly, but recalls having acted altruistically, should be more likely to recall a good outcome. Figure 3 shows the data split along this dimension, with correct recall of the action's selfishness defined as previously. We find that being incorrect about the action's selfishness collapses the rate of outcome recall of non-selfish dictators, but leaves it almost unchanged for selfish dictators. Columns (3) and (4) of Table 3 show the corresponding results in regression form: correct recall of the action's selfishness strongly predicts recall of the outcome for non-selfish dictators ($\beta = 0.311$ in column (3)), but has a large negative interaction with the dictator being selfish ($\beta = -0.267$ in column (3)), so that the resulting effect is small and non-significant for selfish dictators (F-test of the linear combination with the column (3) model: $F(1) = 0.34, p = 0.558$).

As an additional, and more straightforward piece of evidence, we find that among dictators who incorrectly recall their action's selfishness, the selfish dictators who had an

Figure 3: Rate of correct recall of the recipient's outcome among dictators, depending on correct or incorrect recall of having acted selfishly. The dotted line represents the chance level (50%).
Bars represent non-robust standard errors. Stars correspond to χ^2 tests of proportions: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

outcome incongruent with their recalled altruism are significantly more likely to recall it than the non-selfish dictators who had an outcome incongruent with their recalled selfishness (χ^2 test of proportion: $\chi^2(1) = 6.02, p = 0.014$), whereas the inference effect should dampen both similarly [6]. Overall, these results do not conform to the predicted inversion for dictators who incorrectly recall their action's selfishness, and thus provide good evidence that Result 2b is not due to the dictators' recalling the most common outcome given their action. Instead, they seem to be best understood through the conjunction of two effects: first, accurate memory of one detail is correlated with accurate memory of the other among non-selfish dictators; second, bad outcomes are especially memorable for selfish dictators.

Are dictators unsure about their memories, or do they falsely recall an outcome? Existing literature distinguishes between memory errors, and false memories (Chew et al., 2020). To shed light on this question, we examine the probabilistic beliefs that the dictators reported regarding what the recipient's outcome was. Figure 4 shows the distribution of these reported beliefs depending on the dictator's action and the recipient's outcome. Descriptively, we find that 35.2% of dictators report a 100% probability for the correct outcome, with little variation across groups (range: 31.1% – 39.7%), while only 2.6% of dictators report a 0% probability for the correct outcome (8 participants in total). Thus, there are few truly false memories, and the main differences in memory accuracy

[6] Importantly, the null results on differences in the recall of the action's selfishness exclude the possibility that selfish dictators who recall acting altruistically do so motivatedly because they recall that the outcome was bad.

across groups come from the majority of dictators who have some uncertainty about what the recipient's outcome was. Note that the preregistered analyses used these reported beliefs as the main dependent variable for memory of the outcome; we refer the reader to Appendix A.1 for the corresponding results, which are qualitatively identical to the results with the binary outcome.

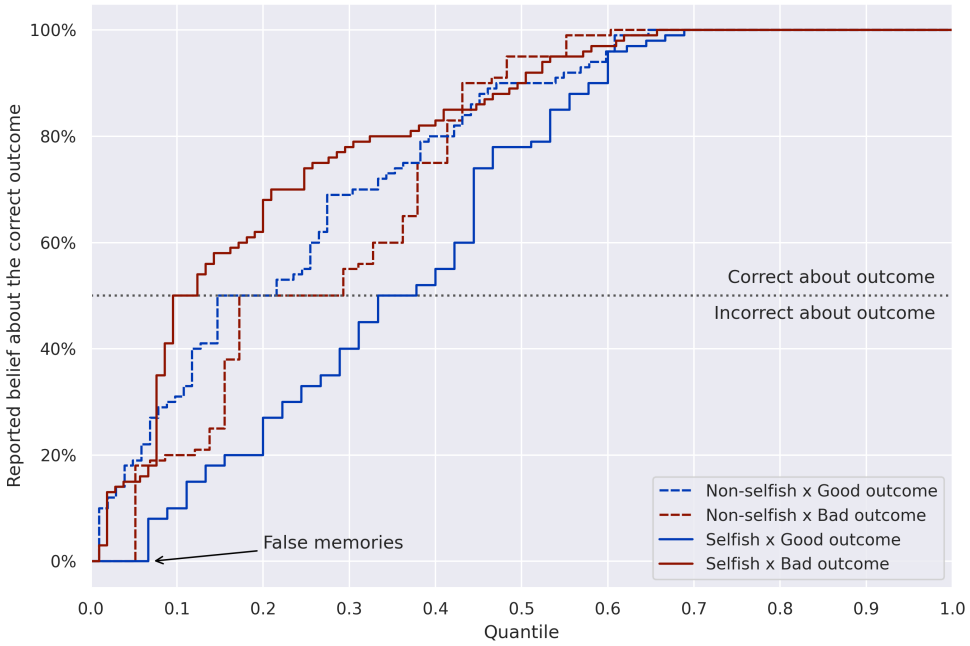


Figure 4: Empirical cumulative distribution functions of the dictators' reported beliefs about the correct outcome (which the recipient actually experienced) depending on the dictator's action and the recipient's outcome. The part of the distribution below the 50% line corresponds to participants incorrectly recalling the outcome.

4.4 Does guilt make bad outcomes more memorable?

In the previous section, we exclude a number of potential mechanisms that could cause dictators who act selfishly to be more likely to correctly recall the recipient's outcome when it was bad. The main constraint was that a valid explanation should account for this pattern, alongside with the absence of difference for dictators who act altruistically. Mechanisms grounded in human moral emotions and cognition, which can attribute the differences in the memories of selfish and non-selfish dictators to the fact that they differ in their moral responsibility in the situation, are thus natural candidates. We now turn our attention to the emotion of guilt, one of the few indisputably moral emotions, as a potential mechanism (Fitouchi et al., 2024).

Guilt is especially relevant in our case for two reasons. First, although, to the best of our knowledge, this hypothesis has yet to be tested, there are both intuitive and theoretical reasons to think that the experience of guilt could depend on outcomes. To build intuition, we asked participants in a pilot experiment to report memories of instances in which

they felt guilty ($N = 45$ participants out of 51 answered this question). Manually coding the answers, we find that 28.9% of respondents report a story in which there was some unexpected realization, providing anecdotal evidence for the idea that guilt is sensitive to outcomes. On the theoretical side, psychological literature has argued that the experience of guilt serves specific functions in encouraging reparative behavior (Baumeister et al., 1994), or learning (Tooby & Cosmides, 2008). Realized outcomes can matter for both of these functions: people sometimes engage in *ex post* redistribution even when the procedure was fair *ex ante* (Cappelen et al., 2013), and general processes of learning under uncertainty rely on integrating realized outcomes as reinforcement.

Of course, the second reason why guilt is especially relevant to our setting is that these adaptive functions of the experience of guilt synergize especially with enhanced recall of the episode, as both reparation and learning require some form of information retention. As a first step in this direction, we test the preregistered hypotheses on the outcome-sensitivity of guilt, H3 and H3'. Table 4 reports the results of regressions and validity measures for each of the four emotions elicited at the end of the initial session: guilt, happiness, annoyance, and pride. To assess the validity of the participants' ratings, we ask them to report again in the follow-up session whether their participation in the initial session made them feel the four emotions. We find that participants answer these questions reliably: guilt, happiness and pride ratings have strong correlations between the initial session and the follow-up session, providing an acceptable level of internal consistency of the measures (Cronbach's $\alpha > 0.767$). Annoyance ratings show a slightly lower correlation and internal consistency ($\alpha_{annoyance} = 0.638$).

We find clear evidence of sensitivity to outcomes in the dictators' guilt ratings, supporting H3: if the recipient experiences a bad outcome, dictators report a rating on average 1.34 points higher (out of 10), corresponding to 43.6% of a standard deviation. In fact, the effect is larger than the effect of acting selfishly, which is not significant, although positive. Finally, the effect of the outcome is significant for both selfish and non-selfish dictators, supporting H3 and not H3'.

Result 3 – In support of H3, but not H3', we find that dictators report experiencing more guilt in the initial session after the recipient lost his endowment, whether they acted selfishly or not.

It may seem surprising that dictators who did not act selfishly still express guilt over the recipient's outcome. We suggest two interpretations. First, dictators may always attribute the responsibility of the outcome to themselves, leading to erroneous attributions among non-selfish dictators, which would be reminiscent of the phenomenon of survivor's guilt. However, this would be rather surprising given the low stakes and clear structure of the environment, which in no way match the usually extreme conditions that foster

Table 4: Self-reported emotion levels in the initial session (scales from 0 to 10)

	Guilt	Happiness	Annoyance	Pride
Constant	1.417***	6.704***	1.833***	5.620***
	(0.246)	(0.218)	(0.243)	(0.298)
Selfish choice (S)	0.801	0.231	0.058	-1.338*
	(0.502)	(0.411)	(0.490)	(0.525)
Bad outcome	1.490**	-0.485	0.745	-1.230*
	(0.470)	(0.393)	(0.455)	(0.482)
Bad outcome × S	-0.322	0.296	-0.300	1.857**
	(0.701)	(0.573)	(0.677)	(0.711)
Adjusted R^2	0.066	-0.001	0.002	0.018
Observations	340	340	340	340
Correlation with follow-up	0.626	0.622	0.469	0.629
Test-retest reliability (α)	0.770	0.767	0.638	0.773

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Linear regressions of each emotion reported by dictators at the end of the initial session, with heteroskedasticity-consistent standard errors. Correlations are between the emotion level reported at the end of the initial session, and the recalled emotion level as reported in the follow-up session. Test-retest reliability corresponds to the standardized alpha coefficient for these two reports.

survivor's guilt.

Secondly, dictators may refer to guilt in a broader sense than strictly moral guilt. Psychological literature on guilt has identified heterogeneity in the meanings given to the word, and that guilt proneness is highly correlated with empathic tendencies and perspective taking in psychological questionnaires (Tangney & Dearing, 2002, report correlations between 0.3 and 0.5 across many different studies). Thus, some dictators may simply be reporting feeling "guilty" in the sense of "feeling bad for the recipient", a basically empathic concern.

Under either interpretation, the experience of guilt as commonly understood is clearly sensitive to outcomes. Can the outcome-sensitivity of guilt explain why dictators who acted selfishly remember bad outcomes so accurately? At first glance, the fact that non-selfish dictators also report experiencing guilt over bad outcomes, despite not recalling bad outcomes more accurately, may seem to limit guilt's explanatory power. However, note that the adaptive functions of guilt mentioned above, reparation and learning, are meaningful only when the individual is responsible for the harm caused. We thus explore whether guilt might have different effects on the dictators' memories, depending on whether they are responsible for the outcome.

Table 5 reports the results of regressions predicting the dictators' recall of the outcome,

with the guilt ratings from the initial session included as predictor. Across all specifications, we find that there is a significant interaction between reported guilt and responsibility: while non-selfish dictators do not experience improved recall when they felt guilty, selfish dictators do. A one standard deviation increase in reported guilt for selfish dictators has a total effect of increasing the rate of outcome recall from 5.3 to 7.5 percentage points on average, an effect size similar to that of a one standard deviation increase in memory ability (5.1 to 6.1 percentage points). This effect of guilt on recall is significant even when controlling the recipient's outcome, memory ability, and the other emotions (p-values are reported in the last line of the table; all $p < 0.04$). In Appendix A.4, we show that none of the other emotions is predictive of outcome recall, despite there being similar variation in guilt and the other emotions, and despite annoyance showing a moderate correlation with guilt ($r = 0.526$; 95% CI: $[0.444, 0.599]$).

Table 5: Guilt and the dictators' recall of the outcome

	(1)	(2)	(3)	(4)	(5)
Constant	0.758***	0.752***	0.779***	0.774***	0.786***
	(0.034)	(0.034)	(0.041)	(0.041)	(0.115)
Selfish choice (S)	0.013	0.027	-0.164*	-0.152	-0.176
	(0.050)	(0.051)	(0.083)	(0.084)	(0.166)
Bad outcome			-0.058	-0.060	-0.042
			(0.075)	(0.073)	(0.072)
Bad outcome × S			0.291**	0.295**	0.269*
			(0.108)	(0.108)	(0.108)
Reported guilt	-0.042	-0.037	-0.036	-0.030	-0.021
	(0.034)	(0.034)	(0.035)	(0.035)	(0.041)
Reported guilt × S	0.112*	0.112**	0.089*	0.089*	0.096*
	(0.044)	(0.043)	(0.044)	(0.043)	(0.049)
Memory ability		0.061*		0.056*	0.051*
		(0.025)		(0.025)	(0.025)
Other emotions × S	✗	✗	✗	✗	✓
Adjusted R^2	0.015	0.035	0.042	0.063	0.055
Observations	310	306	310	306	306
$\beta_{guilt} + \beta_{guilt \times S} = 0$	0.0104	0.00429	0.04	0.0194	0.00334

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of correct recall of the outcome among dictators, with heteroskedasticity-consistent standard errors. The sample includes dictators who completed the follow-up session within the 24-hour window. Columns (2), (4) and (5) exclude those who reported having cheated on the memory task. Column (5) also controls for happiness, annoyance and pride, as well as their interactions with choosing the selfish action.

Finally, it is important to note that there is still an important direct effect of bad outcomes on the selfish dictators' recall rate, even when we include guilt ratings in the regression. Performing a bootstrap mediation analysis with the R package *mediation* (Tingley et al., 2014), the average direct effect of a bad outcome on memory of the outcome among selfish dictators is an important fraction of the total effect (average direct effect: 0.233, 95% CI $[0.082, 0.389], p = 0.001$; total effect: 0.254, 95% CI $[0.102, 0.407], p < 0.001$), such that reported guilt is not a significant mediator of the effect (proportion mediated: 8.2%, 95% CI $[-0.010, 0.279], p = 0.090$). Of course, we do not contend that guilt drives the entirety of the effect of bad outcomes on recall, only that it is one part of a larger picture which would also include moral cognitive mechanisms, not just affective ones. Indeed, the fact that guilt has an effect on memory only when the individual is responsible for the outcome shows that the emotion is integrated to a broader system of moral cognition (see Fitouchi et al., 2024; Railton, 2014; Cosmides et al., 2018, for ongoing discussions of the integration of cognition and affect in moral psychology).

4.5 Behavioral consequences

So far, we found that bad outcomes can be particularly memorable when the individual is responsible for them, and that this is in part due to the experience of guilt being sensitive to outcomes. We now test whether we can complete the chain by identifying an effect of bad outcomes on downstream behavior. Drawing inspiration from the psychological literature on guilt, we test both for an effect on reparative behavior, and for a learning effect.

Table 6 provides the results of the corresponding exercise. On reparative behavior, the main dependent variable is the share of the \$15 endowment allocated to the former recipient. We find that dictators allocate the endowment very equally between their former recipient and the third-party, with 57.1% allocating between \$7 and \$8 to each, and we do not observe any significant differences depending on the dictator's previous action or its outcome, or of any other recorded variable. On learning, we find that participants understand the multiple-price-list version of the dictator-game variant well, with only 7.1% showing preference reversals. The main dependent variable on this task is the price at which these dictators switch to the selfish option. Again, we do not observe any significant differences depending on the outcome of the previous game. The only effect is that dictators who acted selfishly in the previous game tend to do so again in this new one.

Result 4 – We do not find evidence in favor of H4. Dictators do not allocate the endowment differently in the third-party allocation task after the recipient lost his endowment, or when they have memories of the recipient losing his endowment, neither do they act differently in the follow-up dictator game.

Table 6: Behavioral tasks in the follow-up session

	Share allocated to partner	Switching point to selfish ($)	
	(1)	(2)	(3)
Constant	0.538***	3.431***	3.253***
	(0.016)	(0.165)	(0.205)
Selfish choice (S)	-0.008	-0.871***	-0.446
	(0.037)	(0.234)	(0.378)
Bad outcome	0.024		0.512
	(0.026)		(0.342)
Bad outcome × S	-0.021		-0.792
	(0.045)		(0.502)
Reported guilt	0.010		-0.126
	(0.010)		(0.105)
Adjusted R^2	-0.005	0.044	0.046
Observations	301	278	278

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Linear regressions of the share allocated to their partner by dictators in the third-party allocation task, and of the amount of money which leads them to switch to the selfish action in the dictator-game variant. The 7.1% of participants with multiple switching points are excluded from this second regression.

Of course, these null results do not indicate that there are no downstream behavioral consequences of moral concerns about harm whatsoever. In fact, there is already some evidence that guilt has behavioral consequences which can be at the expense of a third-party (de Hooge et al., 2011). One possible explanation of our results could be that the tasks were not appropriate, for instance because participants were unwilling to create additional harm by making an unequal allocation. Another would be that for such a short social interaction, any behavioral consequence will be quite small or short-lived, while a more high-stakes social interaction would have more lasting behavioral consequences. We view the identification of measurable behavioral consequences resulting from the memory effect identified in this paper as an important topic for further research.

5 Robustness

We now provide a number of robustness checks for our main results.

5.1 Attrition

An important concern for the robustness of our results is the fact that we observe some differential attrition depending on the outcome among selfish dictators, with bad outcomes leading to 9.0 percentage points fewer dictators coming back for the follow-up session. Table 7 reports the corresponding regressions. The effect of bad outcomes on the attrition rate of selfish dictators is significant; the corresponding p-values appear in the last line of the table (all $p < 0.021$).

Table 7: Participation in the follow-up session

	(1)	(2)	(3)
Constant	0.944***	0.943***	0.938***
	(0.022)	(0.022)	(0.022)
Selfish choice (S)	0.034	0.036	0.035
	(0.031)	(0.032)	(0.033)
Bad outcome	-0.038	-0.035	-0.023
	(0.043)	(0.044)	(0.041)
Bad outcome × S	-0.055	-0.056	-0.061
	(0.056)	(0.056)	(0.056)
Reported guilt		-0.006	
		(0.016)	
Reported annoyance			-0.060**
			(0.020)
Adjusted R^2	0.006	0.004	0.052
Observations	340	340	340
$\beta_{Bad} + \beta_{Bad \times S} = 0$	0.0102	0.0115	0.0207

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Linear probability models of the dictators' participation in the follow-up session, with heteroskedasticity-consistent standard errors.

Reassuringly, our main result on recall of the outcome is robust to this issue. If we assumed that there was no attrition whatsoever over bad outcomes for selfish dictators, and that the missing dictators were at chance level on recalling the outcome, the effect would still be significant (χ^2 test of proportions: $\chi^2(1) = 7.16, p = 0.007$). In fact, it would still be significant at the 5% level if only 20% of the dictators missing after a bad outcome would have recalled it correctly. Thus, the result is robust to very conservative assumptions about the missing dictators.

Interestingly, reported guilt is not particularly predictive of attrition, while reported annoyance is (see Table 7). These findings are interesting on their own, as they represent a behavioral consequence of concerns for harm among selfish dictators particularly. Not

showing up to the follow-up session leads to foregoing the bonus payment entirely, while the decision to act selfishly was motivated by this bonus payment in the first place. As a note of caution, we do not find the same attrition effect when restricting the sample to dictators who completed the follow-up within an 8-hour window (see Table A.5).

5.2 Is the outcome recall effect due to inattention?

An important question is whether the effect observed really is a memory effect, or if it can be attributed to other factors, such as inattention. Of course, an attention-grabbing event will leave a stronger mark in memory, so the two are not completely opposed. However, if participants did not encode any information about the outcome, then it would be wrong to attribute the effect to memory.

To test whether participants really encode the information, we ran a smaller experiment with $N = 62$ participants (31 dictators), in which participants completed the two experimental sessions in a row. Between the two sessions, participants filled questionnaires on unrelated topics (artificial intelligence and moral universalism), which took a few minutes to be filled. One important difference with the main experiment is that recall was not incentivized.

After the filler questionnaire, we find that 93.5% of dictators and 90.3% of recipients correctly recalled what the outcome was. The corresponding Wilson 95% confidence interval for the proportion of participants correctly recalling the outcome overall is [82.5%, 96.5%]. Thus, participants really encode the information about the outcome, and the effect which we observe in the main experiment is due to different retention dynamics in memory.

6 Discussion

This paper studies how two dimensions of moral preferences, image concerns and concerns about harm, influence the memory of a social interaction. To do so, we use an experimental setting which decorrelates the selfishness of actions and their outcomes, that is, their consequences on others. Our main finding is that bad outcomes are especially memorable for participants who acted selfishly. This effect cannot be attributed to a number of purely cognitive effects, and we instead attribute it to the properly moral difference that it makes to be responsible for harming someone else. In this regard, we find that the experience of guilt increases recall accuracy when the individual is morally responsible for the outcome, although it does not explain the totality of the effect.

Our findings suggest that a quite refined view of memory in the moral domain is warranted. Recalling or not recalling is not simply a matter of hedonic choice, given that the experience of recalling harming someone is unpleasant and yet very well recalled.

However, the main alternative mechanisms of memory currently integrated to behavioral economics are purely cognitive and are not sensitive to parameters such as the individual's social situation (see e.g. Bordalo et al., 2020; Graeber et al., 2024), and therefore cannot explain our results either. Instead, what is needed is a mechanism which is both non-intentional, yet responsive to the individual's payoffs and social situation.

We propose that drawing from work in evolutionary psychology might help economists. If moral affects and cognition have evolved to solve specific cooperation problems, then thinking about these problems can shed light on the specific cognitive adaptations which were developed to solve them (Tooby & Cosmides, 2008; Fitouchi et al., 2024). Of particular interest to the two dimensions of moral preferences studied here is the dual-system theory of moral judgments put forward by Cushman (2008) (for a broad overview, see Cushman, 2025). Cushman argues that moral judgments are sensitive to both the intentional and causal properties of actions, in part because they reflect solutions to two different problems in the evolution of cooperation. Intentional properties matter for partner choice, that is, choosing who to cooperate with (see e.g. Baumard et al., 2013), while causal properties matter for partner control, that is, enforcing cooperation in existing relationships. The outcome sensitivity of guilt discovered in our study seems to put it squarely on the causal side of the divide, coherent with its role in long-term personal relationships (Baumeister et al., 1994).

This study has a number of limitations. First, the mixed findings on motivated forgetting suggest that some experimental parameters matter in making motivated forgetting possible, but we cannot point to which differences with previous work are relevant, given that several parameters change at the same time. Potential candidates include the anonymity of interactions, whether they are one-shot or repeated, and perhaps the exact method of memory elicitation. Second, the experiment is an online, one-shot interaction, with relatively low stakes. Thus, it might minimize the effects of moral concerns, especially of guilt, given that the reported levels of the emotion are in the absolute quite low. Finally, the lack of measured behavioral consequences, apart from an effect on attrition, leaves the question of exactly how the studied psychological effects translate into economically relevant behavior open.

References

Amelio, A., & Zimmermann, F. (2023). Motivated Memory in Economics—A Review. *Games*, *14*(1), 15. https://doi.org/10.3390/g14010015

Battigalli, P., & Dufwenberg, M. (2007). Guilt in Games. *American Economic Review*, *97*(2), 170–176. https://doi.org/10.1257/aer.97.2.170

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *The Behavioral and Brain Sciences*, *36*(1), 59–78. https://doi.org/10.1017/S0140525X11002202

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*, 243–267. https://doi.org/10.1037/0033-2909.115.2.243

Bénabou, R., Falk, A., Henkel, L., & Tirole, J. (2023). Eliciting Moral Preferences Under Image Concerns: Theory and Evidence. *CRC TR 224 Discussion Paper Series*, (crctr224_2023_441). Retrieved June 5, 2026, from https://ideas.repec.org//p/bon/boncrc/crctr224_2023_441.html

Bénabou, R., & Tirole, J. (2002). Self-Confidence and Personal Motivation. *The Quarterly Journal of Economics*, *117*(3), 871–915. https://doi.org/10.1162/003355302760193913

Bénabou, R., & Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652

Bénabou, R., & Tirole, J. (2011). Identity, Morals, and Taboos: Beliefs as Assets. *The Quarterly Journal of Economics*, *126*(2), 805–855. https://doi.org/10.1093/qje/qjr002

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2019). *299 Pictures for Psycholinguistic Studies*. American Psychological Association. https://doi.org/10.1037/t71202-000

Bordalo, P., Gennaioli, N., & Shleifer, A. (2020). Memory, Attention, and Choice. *The Quarterly Journal of Economics*, *135*(3), 1399–1442. https://doi.org/10.1093/qje/qjaa007

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition*, *5*(1), 73–99. https://doi.org/10.1016/0010-0277(77)90018-X

Cappelen, A. W., Konow, J., Sørensen, E. Ø., & Tungodden, B. (2013). Just Luck: An Experimental Study of Risk-Taking and Fairness. *American Economic Review*, *103*(4), 1398–1413. https://doi.org/10.1257/aer.103.4.1398

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, *11*(1), 2100. https://doi.org/10.1038/s41467-020-15602-4

Celebi, C., Exley, C., Harrs, S., Kivimaki, H., Serra-Garcia, M., & Yusof, J. (2026). *Mission Possible: The Collection of High-Quality Data*. 6405479. Retrieved March 29, 2026, from https://papers.ssrn.com/abstract=6405479

Chew, S. H., Huang, W., & Zhao, X. (2020). Motivated False Memory. *Journal of Political Economy*, *128*(10), 3913–3939. https://doi.org/10.1086/709971

Cosmides, L., Guzmán, R. A., & Tooby, J. (2018). The Evolution of Moral Cognition. In *The Routledge Handbook of Moral Epistemology*. Routledge.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F. (2025). Moral Judgment. In *The Handbook of Social Psychology*. Retrieved June 3, 2026, from https://oar.princeton.edu/handle/88435/pr1862bc1w

de Hooge, I. E., Nelissen, R. M. A., Breugelmans, S. M., & Zeelenberg, M. (2011). What is moral about guilt? Acting "prosocially" at the disadvantage of others. *Journal of Personality and Social Psychology*, *100*(3), 462–473. https://doi.org/10.1037/a0021459

Dikmen, S. S., Bauer, P. J., Weintraub, S., Mungas, D., Slotkin, J., Beaumont, J. L., Gershon, R., Temkin, N. R., & Heaton, R. K. (2014). Measuring Episodic Memory Across the Lifespan: NIH Toolbox Picture Sequence Memory Test. *Journal of the International Neuropsychological Society*, *20*(6), 611–619. https://doi.org/10.1017/S1355617714000460

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences. *Journal of the European Economic Association*, *9*(3), 522–550. https://doi.org/10.1111/j.1542-4774.2011.01015.x

Dostoyevsky, F. (2006, March 28). *Crime and Punishment* (C. Garnett, Trans.). Retrieved February 12, 2026, from https://www.gutenberg.org/ebooks/2554 (Original work published 1866)

Fehr, E., & Fischbacher, U. (2002). Why social preferences matter – the impact of non-selfish motives on competition, cooperation and incentives. *Econ. J. (London)*, *112*(478), C1–C33.

Fitouchi, L., André, J.-B., & Baumard, N. (2024). Are There Really So Many Moral Emotions? Carving Morality at Its Functional Joints. In Laith Al-Shawaf & T. K. Shackelford (Eds.), *The Oxford Handbook of Evolution and the Emotions*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197544754.013.35

Galeotti, F., Saucet, C., & Villeval, M. C. (2020). Unethical amnesia responds more to instrumental than to hedonic motives. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(41), 25423. https://doi.org/10.1073/pnas.2011291117

Gödker, K., Jiao, P., & Smeets, P. (2025). Investor Memory. *The Review of Financial Studies*, *38*(6), 1595–1640. https://doi.org/10.1093/rfs/hhaf006

Graeber, T., Roth, C., & Zimmermann, F. (2024). Stories, Statistics, and Memory. *The Quarterly Journal of Economics*, *139*(4), 2181–2225. https://doi.org/10.1093/qje/qjae020

Gurdal, M. Y., Miller, J. B., & Rustichini, A. (2013). Why Blame? *Journal of Political Economy*, *121*(6), 1205–1247. https://doi.org/10.1086/674409

Hao, L., & Houser, D. (2012). Belief elicitation in the presence of naïve respondents: An experimental study. *Journal of Risk and Uncertainty*, *44*(2), 161–180. https://doi.org/10.1007/s11166-011-9133-1

Huffman, D., Raymond, C., & Shvets, J. (2022). Persistent Overconfidence and Biased Memory: Evidence from Managers. *The American Economic Review*, *112*(10), 3141–3175. https://doi.org/10.1257/aer.20190668

Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica*, *77*(2), 603–606. https://doi.org/10.3982/ECTA7833

Kensinger, E. A. (2007). Negative Emotion Enhances Memory Accuracy: Behavioral and Neuroimaging Evidence. *Current Directions in Psychological Science*, *16*(4), 213–218. https://doi.org/10.1111/j.1467-8721.2007.00506.x

Kouchaki, M., & Gino, F. (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences*, *113*(22), 6166–6171. https://doi.org/10.1073/pnas.1523586113

LaForte, E. M., Hook, J. N., & Giella, A. K. (2024). National Institutes of Health (NIH) Toolbox® V3 Technical Manual. *Toolbox Assessments. Inc.* https://nihbabytoolbox.org/wp-content/uploads/2024/04/NIHTB-V3-Technical-Manual_040524.pdf

Levine, L. J., & Pizarro, D. A. (2004). Emotion and memory research: A grumpy overview. *Social Cognition*, *22*(5), 530–554. https://doi.org/10.1521/soco.22.5.530.50767

Li, K. K. (2013). Asymmetric memory recall of positive and negative events in social interactions. *Experimental Economics*, *16*(3), 248–262. https://doi.org/10.1007/s10683-012-9325-9

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2014). Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin & Review*, *21*(3), 629–636. https://doi.org/10.3758/s13423-013-0542-9

McGaugh, J. L. (2013). Making lasting memories: Remembering the significant. *Proceedings of the National Academy of Sciences*, *110*, 10402–10407. https://doi.org/10.1073/pnas.1301209110

Mueller, M. W. (2022). *Essays on Big Life Decisions*. Retrieved June 18, 2026, from https://escholarship.org/uc/item/45r2424p

Railton, P. (2014). The Affective Dog and Its Rational Tale: Intuition and Attunement. *Ethics*, *124*(4), 813–859. https://doi.org/10.1086/675876

Saucet, C., & Villeval, M. C. (2019). Motivated memory in dictator games. *Games and Economic Behavior*, *117*, 250–275. https://doi.org/10.1016/j.geb.2019.05.011

Sial, A. Y., Sydnor, J. R., & Taubinsky, D. (2023, January). *Biased Memory and Perceptions of Self-Control*. 30825. https://doi.org/10.3386/w30825

Smith, A. (2022). *The Theory of Moral Sentiments*. Retrieved June 17, 2026, from https://www.gutenberg.org/ebooks/67363 (Original work published 1759)

Stanley, M. L., Yang, B. W., & De Brigard, F. (2018). No evidence for unethical amnesia for imagined actions: A failed replication and extension. *Memory & Cognition*, *46*(5), 787–795. https://doi.org/10.3758/s13421-018-0803-y

Sznycer, D. (2019). Forms and Functions of the Self-Conscious Emotions. *Trends in Cognitive Sciences*, *23*(2), 143–157. https://doi.org/10.1016/j.tics.2018.11.007

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences*, *113*(10), 2625–2630. https://doi.org/10.1073/pnas.1514699113

Tangney, J. P., & Dearing, R. L. (2002). *Shame and guilt*. Guilford Press. https://doi.org/10.4135/9781412950664.n388

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral Emotions and Moral Behavior. *Annual Review of Psychology*, *58*, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, *59*, 1–38. https://doi.org/10.18637/jss.v059.i05

Tooby, J., & Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In *Handbook of emotions, 3rd ed* (pp. 114–137). The Guilford Press.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Zimmermann, F. (2020). The Dynamics of Motivated Beliefs. *American Economic Review*, *110*(2), 337–361. https://doi.org/10.1257/aer.20180728

A Supplementary analyses

A.1 Preregistered analyses

Deviations with the preregistration. Hypotheses were renumbered for exposition: H1 in the main text corresponds to H3 in the preregistration, H2 and H2' in the main text to H2 and H2' in the preregistration, H3 and H3' in the main text to H1 and H1' in the preregistration, and H4 in the main text to H4 and H5 in the preregistration. We report on hypothesis H6 from the preregistration in the Appendix dedicated to recipients (Appendix A.5).

Regarding the sample size, the preregistration stated the following: "700 participants will be invited to take part in the first session on Prolific, with the aim of having at least 300 dictators take part in the second session (based on the attrition rate of around 10% observed in two pilot experiments). If less than 300 dictators take part in the second session, an additional wave of data collection will be completed to reach this target (overshooting the recruitment in the first session by the observed attrition rate in the initial wave)." However, recruiting 700 participants for the experiment on Prolific within a short time window proved harder than expected, so the recruitment was split in two waves. In total, we recruited 680 participants for the initial session. We stopped at this number because the final goal of having more than 300 dictators completing the follow-up session was already met.

The target of 300 dictators is met for the sample of dictators who completed the follow-up within a 24-hour window ($N_2^{24h} = 310$), but not for dictators who completed the follow-up within an 8-hour window ($N_2^{8h} = 274$), leading us to report the main results with the 24-hour sample, and robustness checks restricting to the 8-hour sample in Appendix A.2. The preregistration stated that we would report the results for both samples. The only qualitatively different results are on attrition (see Section 5.1), and are flagged accordingly.

In the preregistration, we state that the main dependent variable for the memory of the outcome will be the reported subjective beliefs about the outcome, while the binary correctness will be used for robustness. In the main text, we do the opposite for simplicity of exposition, reporting the analyses with reported subjective beliefs in this appendix. The results are qualitatively unchanged.

Result of the preregistered analyses. We report the results of the preregistered analyses, which are qualitatively similar to the analyses reported in the main text. We report on the hypotheses as numbered in the preregistration.

Hypothesis PR1 – Conditioning on the choice, self-reported guilt should be higher when the realized outcome is worse for the recipient. *(outcome-based guilt)*

Result PR1 – Performing an unequal variances t-test on the dictators' reported guilt, we find that self-reported guilt is higher after a bad outcome: $t(337.54) = 4.90, p < 0.001$. We find that it is the case both among selfish dictators ($t(84.60) = 2.22, p = 0.029$) and among non-selfish dictators ($t(110.23) = 3.15, p = 0.002$), supporting H1 but not H1'.

Hypothesis PR2 – Conditioning on the choice, episodic recall should be higher when the realized outcome is worse for the recipient. *(guilt-induced recall)*

Result PR2 – For the memory of the action, the preregistered variable is exact recall of the number of bad balls in the urn. Performing a chi-square test of proportions, we find that the recall after a bad outcome is not significantly higher than after a good outcome ($\chi^2(1) = 0.01, p = 0.935$). This is also the case specifically among selfish dictators ($\chi^2(1) = 0.09, p = 0.768$). For the memory of the outcome, the preregistered variable is the reported subjective belief about the correct outcome. Performing a unequal variances t-test, we find that the reported belief about the correct outcome is not significantly higher after a bad outcome among all dictators ($t(297.43) = -1.46, p = 0.146$), but that it is so among selfish dictators ($t(63.42) = -2.64, p = 0.01$). Thus, we find support for H2', but not H2, on the memory of the outcome, and nothing on the memory of the action.

Hypothesis PR3 – Conditioning on the outcome, episodic recall should be lower when the chosen risk level was higher. *(image-driven motivated recall)*

Result PR3 – For the memory of the action, we preregistered to fit a logistic regression on the subsample of dictators whose recipients did not get the bad outcome, including the chosen action and the measured episodic memory ability as independent variables. Having chosen the selfish action is not a significant predictor of memory of the action ($\beta = 0.342(0.377)$; likelihood-ratio test: $\chi^2(1) = 0.833, p = 0.361$); see Table 2 for these results in the main text. For the memory of the outcome, we preregistered to fit a beta regression of the the reported subjective belief of the correct outcome on the same sample, including the chosen action and the measured episodic memory ability as independent variables. Having chosen the selfish action is not a significant predictor of memory of the action ($\beta = -0.298(0.226)$; likelihood-ratio test: $\chi^2(1) = 1.69, p = 0.194$). Thus, we do not find support for H3 in our preregistered specifications. For the outcome, some of the specifications reported in the main text find a significant effect, so we highlight when relevant the fact the the preregistered specification does not (see Table 3 in the main text).

Hypothesis PR4 – Dictators allocate more to the recipient when the realized outcome was worse for the recipient, and when they recall that it was worse.

Result PR4 – We fit linear regressions on the sample of selfish dictators. We regress the amount allocated to the recipient, depending on the outcome, or on the dictator's reported belief about the outcome. We do not find a significant effect of outcome ($\beta = 0.105\$(0.522)$; likelihood-ratio test: $\chi^2(1) = 0.04, p = 0.840$), or of the memory of the outcome ($\beta = -0.545\$(0.669)$; likelihood-ratio test: $\chi^2(1) = 0.67, p = 0.413$). See Table 6 for complete regression results.

Hypothesis PR5 – Dictators behave more prosocially towards others when the realized outcome was worse for the recipient, and when they recall that it was worse.

Result PR5 – We fit linear regressions on the sample of selfish dictators. We regress the switching point in the multiple-price list, depending on the outcome, or on the dictator's reported belief about the outcome. We do not find a significant effect of outcome ($\beta = -0.332\$(0.367)$; likelihood-ratio test: $\chi^2(1) = 0.83, p = 0.363$), or of the memory of the outcome ($\beta = -0.165\$(0.465)$; likelihood-ratio test: $\chi^2(1) = 0.12, p = 0.721$). See Table 6 for complete regression results.

A.2 Restricted sample (completion within 8 hours)

In this section, we report replications of the main regressions on the restricted sample including only the dictators who completed the follow-up session within the 8-hour window.

Table A.1: Recall of the action among dictators

	Exact recall		Selfishness recall	
	(1)	(2)	(3)	(4)
Constant	0.571***	0.570***	0.714***	0.714***
	(0.052)	(0.052)	(0.047)	(0.047)
Selfish choice (S)	0.054	0.054	0.011	0.011
	(0.092)	(0.096)	(0.085)	(0.089)
Bad outcome	-0.050	-0.053	-0.062	-0.065
	(0.090)	(0.090)	(0.085)	(0.084)
Bad outcome × S	0.043	0.060	0.100	0.109
	(0.128)	(0.130)	(0.118)	(0.120)
Memory ability		0.030		0.022
		(0.031)		(0.029)
Adjusted R^2	-0.005	-0.005	-0.004	-0.005
Observations	274	270	274	270

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of exact recall of the action and correct recall of selfishness among dictators, with heteroskedasticity-consistent standard errors. Exact recall of selfishness is defined as correctly recalling the exact number of red balls in the urn. Correct recall of selfishness is defined as correctly recalling what color was the majority of balls in the urn. The sample includes dictators who completed the follow-up session within the 8-hour window. Columns (2) and (4) excludes those who reported having cheated on the memory task.

Table A.2: Correct recall of the outcome among dictators

	(1)	(2)	(3)	(4)
Constant	0.780***	0.779***	0.556***	0.559***
	(0.043)	(0.043)	(0.083)	(0.083)
Selfish choice (S)	-0.155	-0.155	0.040	0.026
	(0.088)	(0.090)	(0.127)	(0.129)
Bad outcome	-0.041	-0.045	-0.022	-0.025
	(0.078)	(0.078)	(0.076)	(0.076)
Bad outcome × S	0.292*	0.306**	0.271*	0.286*
	(0.114)	(0.115)	(0.113)	(0.114)
Correct about selfishness			0.314***	0.308***
			(0.085)	(0.086)
Correct about selfishness × S			-0.274*	-0.258*
			(0.117)	(0.119)
Memory ability		0.032		0.023
		(0.028)		(0.027)
Adjusted R^2	0.031	0.038	0.088	0.093
Observations	274	270	274	270

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of correct recall of the outcome among dictators, with heteroskedasticity-consistent standard errors. The sample includes dictators who completed the follow-up session within the 8-hour window. Columns (2) and (4) exclude those who reported having cheated on the memory task.

Table A.3: Guilt and the dictators' recall of the outcome

	(1)	(2)	(3)	(4)	(5)
Constant	0.767***	0.764***	0.776***	0.775***	0.728***
	(0.036)	(0.036)	(0.044)	(0.044)	(0.126)
Selfish choice (S)	0.006	0.013	-0.158	-0.158	-0.086
	(0.053)	(0.054)	(0.087)	(0.089)	(0.176)
Bad outcome			-0.026	-0.031	0.005
			(0.081)	(0.080)	(0.078)
Bad outcome × S			0.254*	0.266*	0.227*
			(0.115)	(0.115)	(0.114)
Reported guilt	-0.033	-0.032	-0.030	-0.029	-0.026
	(0.035)	(0.035)	(0.037)	(0.036)	(0.043)
Reported guilt × S	0.106*	0.109*	0.087	0.090*	0.110*
	(0.044)	(0.044)	(0.045)	(0.044)	(0.051)
Memory ability		0.036		0.031	0.021
		(0.028)		(0.028)	(0.028)
Other emotions × S	✗	✗	✗	✗	✓
Adjusted R^2	0.014	0.020	0.038	0.047	0.045
Observations	274	270	274	270	270
$\beta_{guilt} + \beta_{guilt \times S} = 0$	0.00609	0.00313	0.0254	0.0152	0.00157

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Replication of Table 5 with the participants who completed the follow-up session within the 8-hour window.Linear probability models of correct recall of the outcome among dictators, with heteroskedasticity-consistent standard errors. The sample includes dictators who completed the follow-up session within the 8-hour window. Columns (2), (4) and (5) exclude those who reported having cheated on the memory task. Column (5) also controls for happiness, annoyance and pride, as well as their interactions with choosing the selfish action.

Table A.4: Behavioral tasks in the follow-up session

	Share allocated to partner	Switching point to selfish ($)	
	(1)	(2)	(3)
Constant	0.530***	3.376***	3.334***
	(0.017)	(0.179)	(0.227)
Selfish choice (S)	-0.021	-0.839***	-0.438
	(0.040)	(0.252)	(0.404)
Bad outcome	0.036		0.427
	(0.030)		(0.382)
Bad outcome × S	-0.025		-0.722
	(0.050)		(0.547)
Reported guilt	0.005		-0.050
	(0.004)		(0.040)
Adjusted R^2	0.002	0.039	0.039
Observations	269	248	248

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Replication of Table 6 with the participants who completed the follow-up session within the 8-hour window. Linear regressions of the share allocated to their partner by dictators in the third-party allocation task, and of the amount of money which leads them to switch to the selfish action in the dictator-game variant. The 7.1% of participants with multiple switching points are excluded from this second regression.

Table A.5: Participation in the follow-up session within 8 hours

	(1)	(2)	(3)
Constant	0.843***	0.845***	0.887***
	(0.035)	(0.037)	(0.038)
Selfish choice (S)	0.027	0.028	0.028
	(0.061)	(0.061)	(0.060)
Bad outcome	-0.124	-0.122	-0.106
	(0.066)	(0.067)	(0.064)
Bad outcome × S	0.058	0.057	0.050
	(0.090)	(0.090)	(0.088)
Reported guilt		-0.001	
		(0.007)	
Reported annoyance			-0.024**
			(0.008)
Adjusted R^2	0.007	0.004	0.035
Observations	340	340	340
$\beta_{Bad} + \beta_{Bad \times S} = 0$	0.281	0.297	0.36

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Replication of Table 7 for participation within 8 hours. Linear probability models of the dictators' participation in the follow-up session, with heteroskedasticity-consistent standard errors.

A.3 Memory ability measures

Initial session measure. In the initial session, we measure the participants' episodic memory ability with an adaptated version of the standard Picture Sequence Memory Test from the NIH Toolbox Cognition Battery (Dikmen et al., 2014). In the original test, participants are presented a series of fifteen pictures of activities in a sequence, and later have to drag and drop to reorder them according to that sequence. The test measures episodic memory, and not short-term working memory, because the list is too long to be kept in working memory.

For the adapted version, we use fifteen pictures of everyday objects from a set of normed drawings Bonin et al. (2019). Objects are presented with their names. We pick the fifteen best pictures as ranked in terms of image agreement, name agreement, concept familiarity, and visual simplicity. We adapt version 3 of the test, in which participants do only two learning and retrieval rounds on the main set of pictures (LaForte et al., 2024).

Participants first complete a practice round with four other pictures, and then perform the main learning and retrieval rounds on the same set of fifteen pictures. Because we use simple pictures, and in order to make the task more appropriate for an online setting, we decrease slightly the exposure time for pictures (4000 milliseconds per picture instead of 5000). The memory ability measure is the cumulative number of exactly adjacent pairs of pictures remembered correctly over the two main trials.

We take several measures to limit any potential cheating on the task. First, the task is not incentivized, and participants are informed of this fact. Second, we ask participants not to cheat. Third, after having completed the task, participants are asked to report whether they cheated on the task. We clearly state that we will let them complete the rest of the experiment and accept their submission even if they report having cheated. Overall, 1.9% of participants declare having cheated, and we exclude them from the analyses which involve this memory measure.

Follow-up session measure. The follow-up session includes a secondary, custom measure of episodic memory ability, which is derived from the first. Participants are successively exposed to the fifteen pictures from the picture sequence memory test, randomly interleaved with fifteen decoy pictures. They have to press the space bar in order to identify the pictures which were in the initial set. Each picture is displayed for 2500 milliseconds, with 500-millisecond buffers in between.

The score of a participant on this measure is the number of pictures correctly identified as belonging, or not belonging, to the initial set. We find that the scores among dictators do not differ depending on the dictator's action or the recipient's outcome, and are somewhat predictive of correct recall of the action and of the outcome (see Table A.6).

Table A.6: Results with the follow-up measure of episodic memory ability

	Follow-up measure	Action recall	Outcome recall	
	(1)	(2)	(3)	(4)
Constant	0.280***	0.518***	0.759***	0.743***
	(0.059)	(0.051)	(0.044)	(0.047)
Selfish choice (S)	-0.041	0.124	-0.159	-0.138
	(0.106)	(0.092)	(0.087)	(0.086)
Bad outcome	0.015	-0.066	-0.069	-0.043
	(0.107)	(0.084)	(0.075)	(0.077)
Bad outcome × S	-0.081	0.039	0.345**	0.294**
	(0.158)	(0.124)	(0.112)	(0.112)
Guilt				-0.050
				(0.040)
Guilt × S				0.115*
				(0.050)
Initial measure	0.258***			
	(0.043)			
Follow-up measure		0.152***	0.096*	0.086*
		(0.042)	(0.039)	(0.040)
Adjusted R^2	0.131	0.044	0.062	0.073
Observations	274	274	274	274

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
(1): linear regression of the score on the follow-up measure of memory ability among dictators. (2): linear probability model of exact recall of the action among dictators. (3) & (4): linear probability models of correct recall of the outcome among dictators. All models with with heteroskedasticity-consistent standard errors, and including dictators who completed the follow-up session within the 24-hour window.

A.4 Do other emotions predict recall?

In this section, we report on the relationships of the four reported emotions (happiness, annoyance, pride and guilt) with the memories of the action and of the outcome. Table A.7 reports the results for the memory of the action, and Table A.8 for the memory of the outcome. We find that reported guilt among selfish dictators is the only case of an emotion which is robustly associated with a measure of memory, specifically with the memory of the outcome. Reported annoyance is a negative predictor of the memory of the action among selfish dictators, but is significant at the 5% level only when controlling for the other emotions.

Table A.7: Emotions and the dictators' recall of the action

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Constant	0.552***	0.553***	0.549***	0.554***	0.557***	0.558***	0.563***	0.560***	0.559***
	(0.049)	(0.049)	(0.049)	(0.050)	(0.049)	(0.049)	(0.050)	(0.051)	(0.052)
Selfish choice (S)	0.097	0.094	0.096	0.090	0.083	0.082	0.084	0.088	0.096
	(0.086)	(0.086)	(0.087)	(0.086)	(0.087)	(0.087)	(0.088)	(0.090)	(0.090)
Bad outcome	-0.044	-0.050	-0.031	-0.043	-0.045	-0.047	-0.052	-0.047	-0.061
	(0.081)	(0.082)	(0.082)	(0.082)	(0.083)	(0.084)	(0.083)	(0.085)	(0.085)
Bad outcome × S	0.015	0.023	0.006	0.020	0.023	0.024	0.012	0.003	0.002
	(0.118)	(0.118)	(0.119)	(0.118)	(0.119)	(0.120)	(0.119)	(0.124)	(0.123)
Happiness	-0.051	-0.076							-0.077
	(0.027)	(0.040)							(0.052)
Happiness × S		0.048							0.051
		(0.055)							(0.069)
Annoyance			-0.004	0.050					0.036
			(0.028)	(0.042)					(0.050)
Annoyance × S				-0.103					-0.136*
				(0.058)					(0.066)
Pride					-0.030	-0.035			0.007
					(0.028)	(0.040)			(0.051)
Pride × S						0.010			0.011
						(0.057)			(0.072)
Guilt							0.041	0.030	0.012
							(0.029)	(0.043)	(0.050)
Guilt × S								0.021	0.084
								(0.058)	(0.067)
$\beta_{emotion} + \beta_{emotion \times S} = 0$		0.469		0.173		0.532		0.197	
Adjusted R^2	0.007	0.006	-0.004	0.004	-0.000	-0.004	0.002	-0.001	0.010
Observations	310	310	310	310	310	310	310	310	310

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of exact recall of the action among dictators, with heteroskedasticity-consistent standard errors. The sample includes dictators who completed the follow-up session within the 24-hour window.

Table A.8: Emotions and the dictators' recall of the outcome

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Constant	0.784***	0.784***	0.783***	0.782***	0.776***	0.774***	0.789***	0.771***	0.757***
	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.042)	(0.042)	(0.044)	(0.045)
Selfish choice (S)	-0.162	-0.162	-0.161	-0.160	-0.149	-0.148	-0.166*	-0.144	-0.121
	(0.083)	(0.083)	(0.083)	(0.083)	(0.083)	(0.084)	(0.083)	(0.083)	(0.084)
Bad outcome	-0.077	-0.077	-0.074	-0.073	-0.064	-0.059	-0.084	-0.058	-0.039
	(0.072)	(0.072)	(0.073)	(0.073)	(0.072)	(0.072)	(0.073)	(0.075)	(0.073)
Bad outcome × S	0.331**	0.331**	0.329**	0.327**	0.315**	0.311**	0.333**	0.291**	0.259*
	(0.107)	(0.107)	(0.107)	(0.108)	(0.107)	(0.107)	(0.107)	(0.108)	(0.108)
Happiness	0.000	0.001							-0.041
	(0.023)	(0.034)							(0.047)
Happiness × S		-0.001							0.024
		(0.046)							(0.061)
Annoyance			-0.016	-0.022					-0.020
			(0.026)	(0.039)					(0.047)
Annoyance × S				0.011					-0.039
				(0.052)					(0.059)
Pride					0.030	0.041			0.068
					(0.023)	(0.035)			(0.045)
Pride × S						-0.021			-0.021
						(0.046)			(0.060)
Guilt							0.012	-0.038	-0.027
							(0.023)	(0.038)	(0.045)
Guilt × S								0.096*	0.112*
								(0.047)	(0.052)
$\beta_{emotion} + \beta_{emotion \times S} = 0$		0.984		0.759		0.527		0.04	
Adjusted R^2	0.032	0.029	0.034	0.031	0.037	0.035	0.033	0.042	0.040
Observations	310	310	310	310	310	310	310	310	310

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of correct recall of the outcome among dictators, with heteroskedasticity-consistent standard errors. The sample includes dictators who completed the follow-up session within the 24-hour window.

A.5 Recipients

A.5.1 Differences with the experimental design for dictators

The experimental design for recipients had several small differences with the design for dictators, which we list in this section. In the initial session, recipients were recruited asynchronously, and their recruitment started a few minutes after the recruitment of dictators had started. Instead of choosing an urn, recipients saw the dictator's decision screen, with the chosen option highlighted, for the same amount of time that the dictator took to make the decision (capped at 45 seconds). The rest of the initial session, and especially the outcome screen, was designed to be as similar as possible between dictators and recipients, with a few modifications in wording. In the follow-up session, the pronouns of the memory questions were also slightly modified to make them meaningful for recipients. Otherwise, the design was the same for dictators and recipients.

A.5.2 Results

In this section, we report the results of the experiment on the recipient side. Based on previous work on the receiving end of social interaction, our preregistered hypotheses were that recipients would be more likely to forget about the social interaction after the dictator had acted selfishly, and after they were harmed (see Li, 2013; H6 in the preregistration).

Table A.9 and Table A.10 show the corresponding results for the memory of the dictator's action and for the recipient's outcome, which are not in line with these predictions. In particular, note that recipients recall the dictator's selfishness better when they were selfish. For the memory of the outcome, no strong pattern emerges based only on the dictator's action or the outcome. However, when splitting recipients depending on whether they correctly recall the dictator's action, we find that those who wrongly recall a selfish dictator having been non-selfish are especially likely to be mistaken about the outcome, while memory is similar among those who correctly recall the dictator's action (see also Figure A.1).

In terms of emotions, we observe that recipients express much higher levels of annoyance, and lower levels of happiness, after a bad outcome. Interestingly, they also express annoyance when the dictator acted selfishly, even if it did not cause any harm (see Table A.11). In the follow-up session, this annoyance appears to translate to a form of punishment in the allocation task: recipients allocate less to the dictators if they chose the selfish option, or if the outcome was bad for them in the end, even if the dictator did not choose the selfish option. Thus, the results offer a conceptual replication of Gurdal et al. (2013), who find that people sometimes punish others based on outcomes.

Table A.9: Recall of the action among recipients

	Exact recall		Selfishness recall	
	(1)	(2)	(3)	(4)
Constant	0.386***	0.397***	0.604***	0.593***
	(0.048)	(0.050)	(0.049)	(0.050)
Selfish choice (S)	0.172	0.171	0.210**	0.230**
	(0.090)	(0.088)	(0.077)	(0.076)
Bad outcome	-0.024	-0.023	-0.018	0.008
	(0.080)	(0.081)	(0.081)	(0.082)
Bad outcome × S	-0.091	-0.102	-0.032	-0.065
	(0.120)	(0.119)	(0.109)	(0.108)
Memory ability		0.062*		0.059*
		(0.027)		(0.026)
Adjusted R^2	0.006	0.017	0.030	0.045
Observations	308	302	308	302

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of exact recall of the dictator's action and correct recall of the dictator's selfishness among recipients, with heteroskedasticity-consistent standard errors. Exact recall of selfishness is defined as correctly recalling the exact number of red balls in the urn. Correct recall of selfishness is defined as correctly recalling what color was the majority of balls in the urn. The sample includes recipients who completed the follow-up session within the 24-hour window. Columns (2) and (4) excludes those who reported having cheated on the memory task.

Figure A.1: Rate of correct recall of the outcome among recipients, depending on correct or incorrect recall of the dictator's selfishness. The dotted line represents the chance level (50%).

Bars represent non-robust standard errors.

Table A.10: Correct recall of the outcome among recipients

	(1)	(2)	(3)	(4)
Constant	0.792***	0.792***	0.682***	0.692***
	(0.040)	(0.041)	(0.065)	(0.066)
Selfish choice (S)	-0.164	-0.161	-0.294*	-0.312*
	(0.084)	(0.083)	(0.123)	(0.125)
Bad outcome	0.053	0.052	0.056	0.049
	(0.062)	(0.063)	(0.063)	(0.064)
Bad outcome × S	0.140	0.135	0.152	0.154
	(0.103)	(0.103)	(0.101)	(0.101)
Correct about selfishness			0.182**	0.171*
			(0.066)	(0.067)
Correct about selfishness × S			0.112	0.135
			(0.112)	(0.116)
Memory ability		0.019		0.003
		(0.025)		(0.025)
Adjusted R^2	0.018	0.016	0.080	0.077
Observations	308	302	308	302

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear probability models of correct recall of the outcome among recipients, with heteroskedasticity-consistent standard errors. The sample includes recipients who completed the follow-up session within the 24-hour window. Columns (2) and (4) exclude those who reported having cheated on the memory task.

Table A.11: Self-reported emotion levels in the initial session (scales from 0 to 10)

	Guilt	Happiness	Annoyance	Pride
Constant	1.037***	6.815***	1.972***	5.259***
	(0.191)	(0.229)	(0.236)	(0.299)
Selfish choice (S)	0.398	-0.750	1.636**	-1.172*
	(0.424)	(0.433)	(0.532)	(0.562)
Bad outcome	0.338	-1.768***	3.575***	-1.744***
	(0.364)	(0.491)	(0.528)	(0.515)
Bad outcome × S	-0.601	0.317	-1.437	1.493
	(0.570)	(0.687)	(0.782)	(0.767)
Adjusted R^2	-0.005	0.082	0.203	0.035
Observations	340	340	340	340
Correlation with follow-up	0.371	0.697	0.697	0.662
Test-retest reliability (α)	0.541	0.822	0.822	0.797

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear regressions of each emotion reported by recipients at the end of the initial session, with heteroskedasticity-consistent standard errors. Correlations are between the emotion level reported at the end of the initial session, and the recalled emotion level as reported in the follow-up session. Test-retest reliability corresponds to the standardized alpha coefficient for these two reports.

Table A.12: Recipient behavior in the allocation task in the follow-up session

	(1)	(2)
Constant	0.559***	0.528***
	(0.021)	(0.025)
Selfish choice (S)	-0.145***	-0.122***
	(0.037)	(0.036)
Bad outcome	-0.079*	-0.028
	(0.036)	(0.038)
Bad outcome × S	-0.021	-0.040
	(0.053)	(0.051)
Reported annoyance		-0.052***
		(0.015)
Adjusted R^2	0.166	0.199
Observations	293	293

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear regressions of the share allocated to their partner by recipients in the third-party allocation task.

B Experiment instructions

B.1 Decision screens

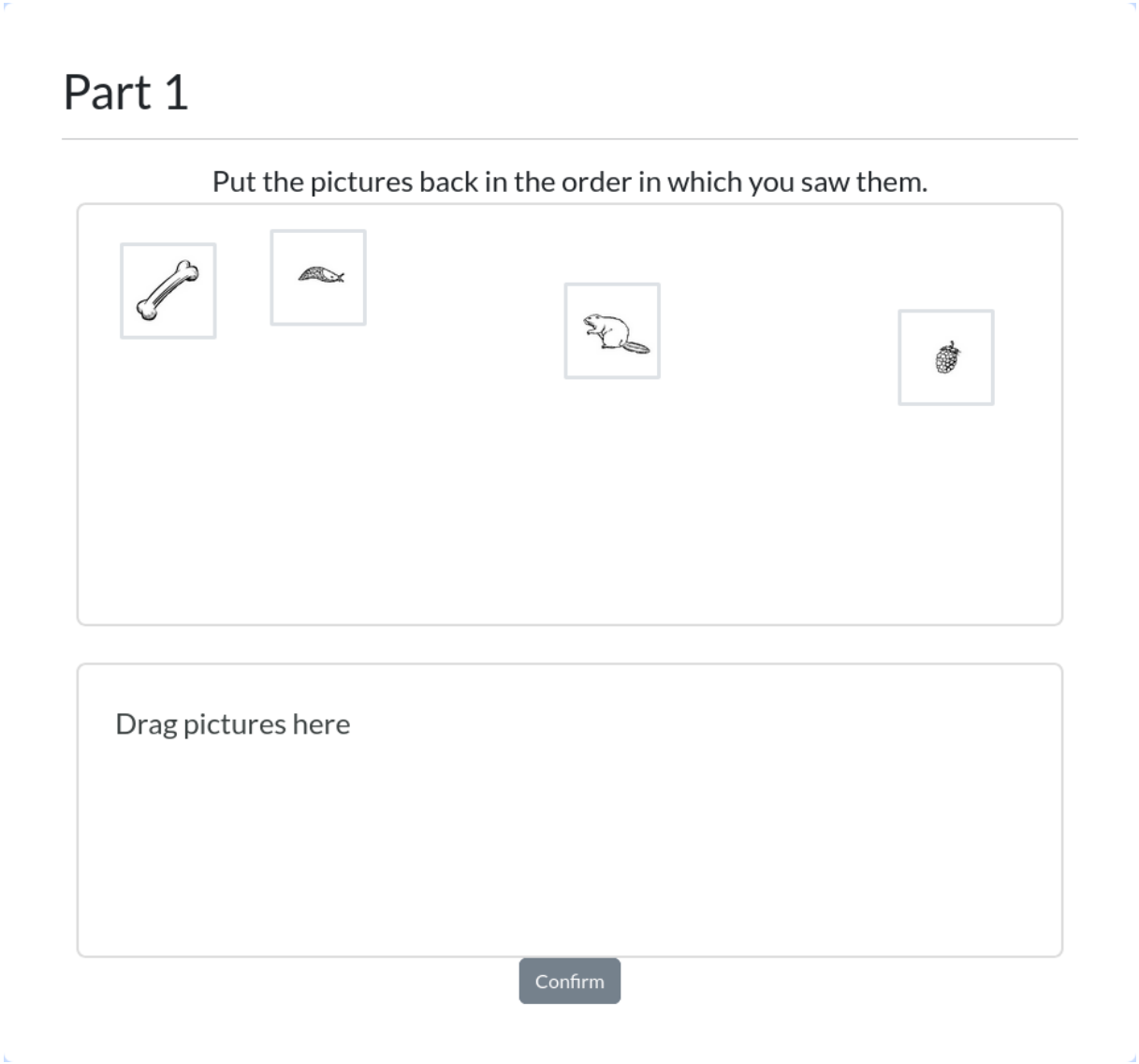


Figure B.1: Practice on the memory ability task in the initial session.

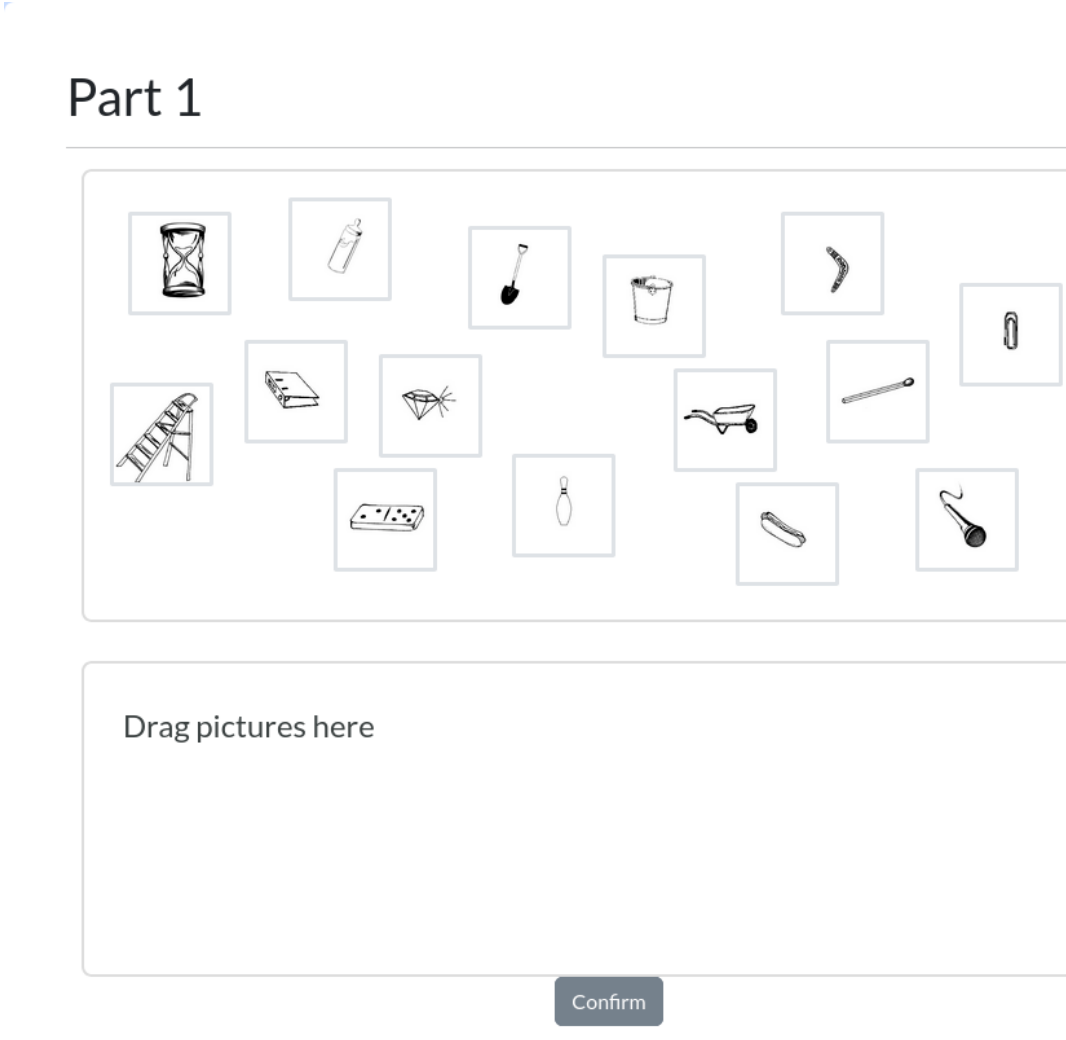


Figure B.2: Memory ability task in the initial session.

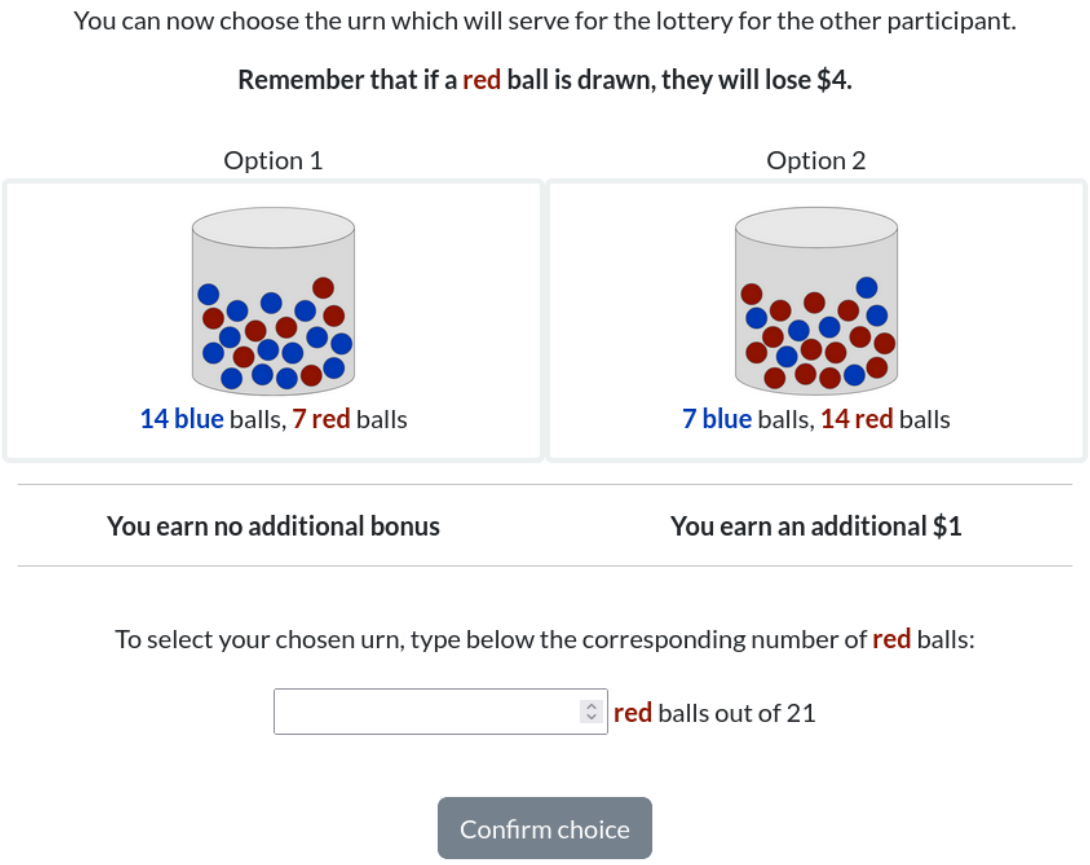


Figure B.3: Decision screen of the dictator.

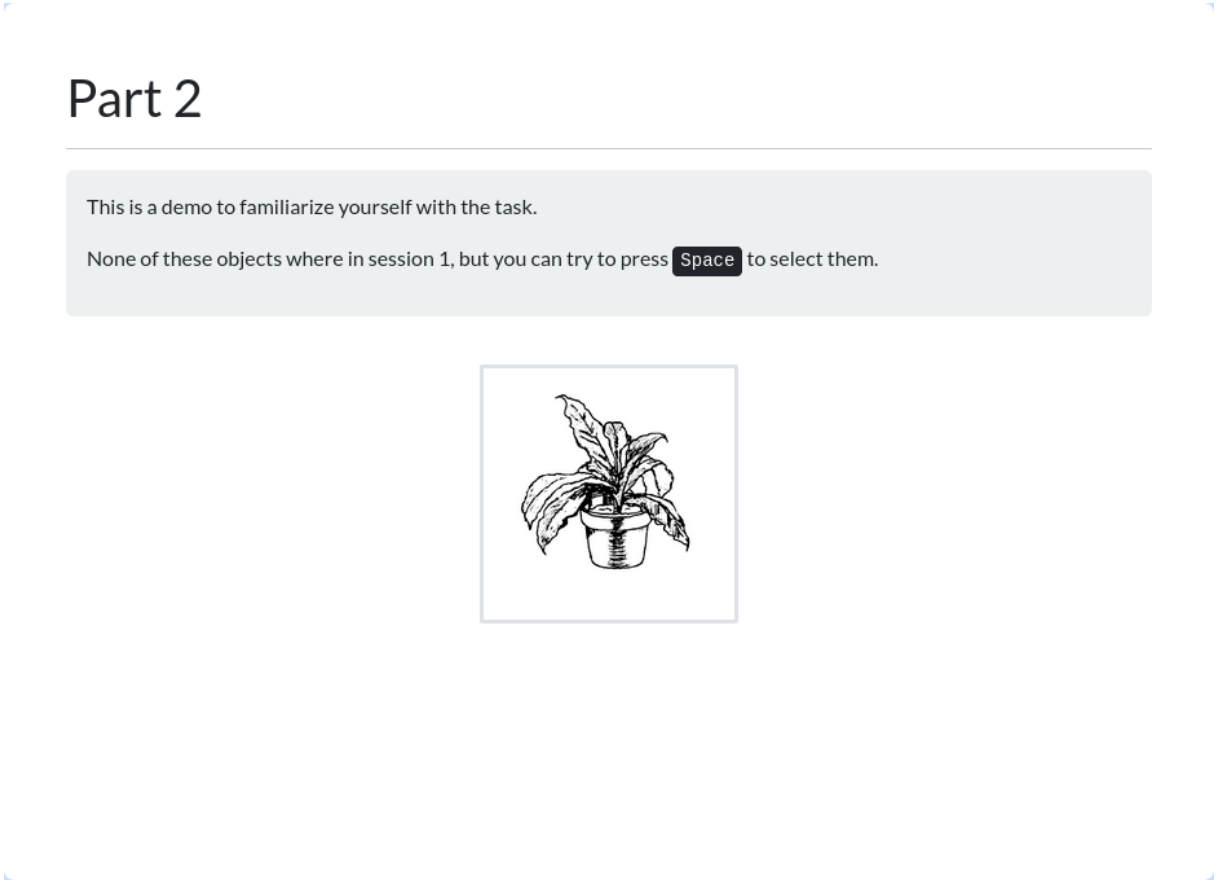


Figure B.4: Practice on the secondary memory ability task in the follow-up session.

Figure B.5: Secondary memory ability task in the follow-up session (selected object).

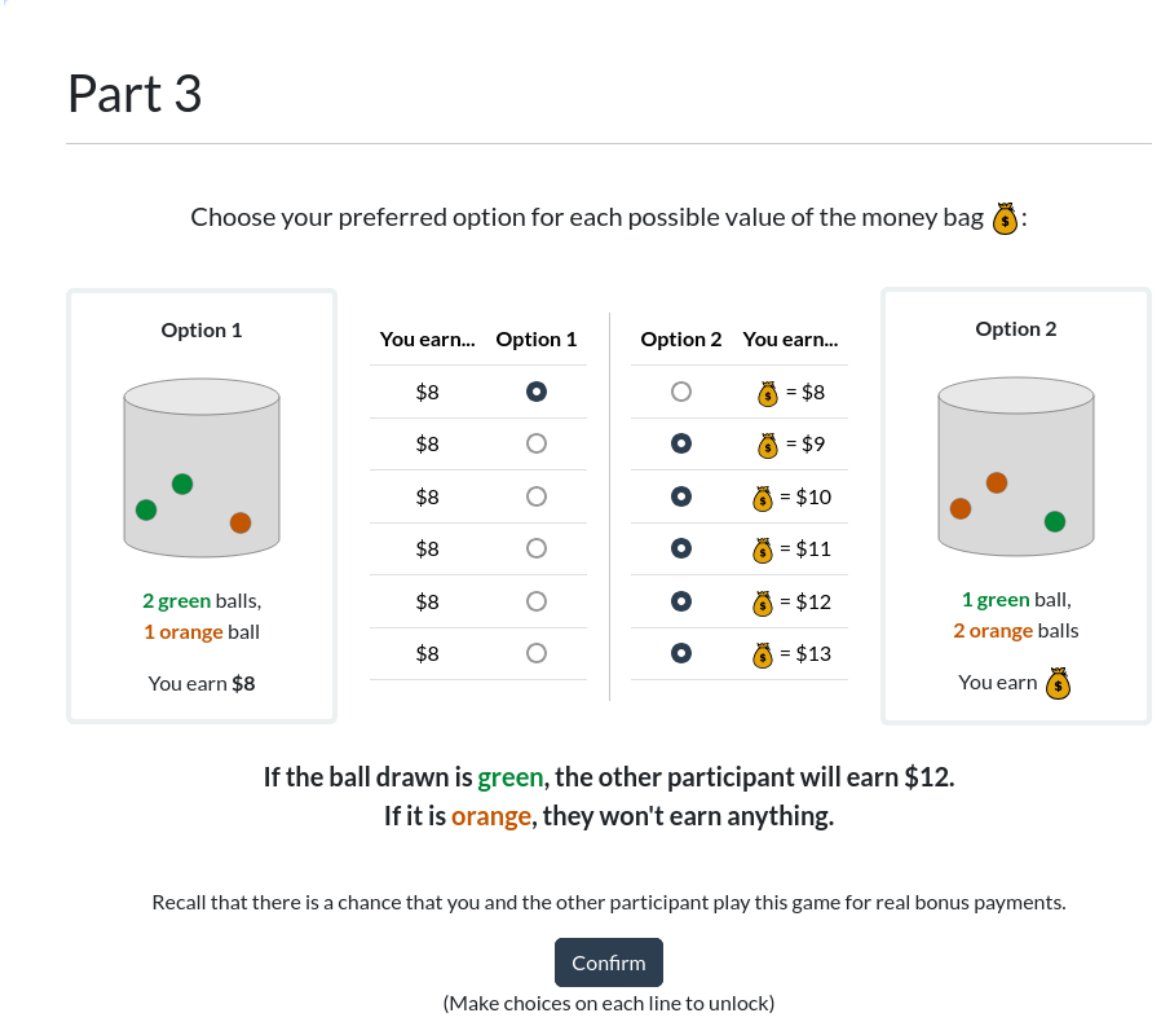


Figure B.6: Multiple-price-list variant of the dictator game.

B.2 Detailed instructions for the main experiment

In this section, we provide the detailed textual instructions for the experiment.

Initial session

Screen 1

Experiment overview

This experiment is divided over two sessions. You are now in the first session, which should last around 10 minutes.

If you complete this session in full, you will be eligible for the second session.

The second session will take place in 48 hours, on Prolific. It should last 8 minutes on average, and will earn you payment at the same base rate.

The second session will be opened for 8 hours, starting on January 29, 2026 at 3:00 PM PDT / 6:00 PM EDT.

We ask you not to complete the first session if you are unsure whether you will be able to complete the second session.

If you are not sure to be able to complete the second session, please click the button below to return your Prolific submission: ⟨Return to Prolific⟩

⟨Next⟩

Screen 2

Bonus payments

In this first session, you might earn up to $5 in bonus payments.

You might gain or lose bonus payments throughout the session, but you will never go below zero.

You will receive your bonus payments upon completion of the second session.

This experiment follows the guidelines of our ethics committee, which does not allow misleading instructions, and more generally deceiving the participant in any way.

⟨Previous⟩ ⟨Next⟩

Screen 3

Browser setting

For this experiment, it is important that your browser fills up the whole screen.

Please press F11 (or Fn + F11) on Windows, or Cmd + Maj + F on Mac, in order to make the browser fullscreen.

You can leave this fullscreen mode at the end of the experiment by pressing Esc.

Warning: Please do not click the back arrow of your web browser during this experiment. If you do this we will have to terminate your participation and ask you to return your submission.

Thank you for your participation!
⟨Previous⟩ ⟨Start⟩

Screen 4
Part 1
In Part 1, we will test your learning capabilities.
First, you will see various pictures displayed on your screen one after the other. Then, you will have to put them back in the order in which you saw them. You can take as long as you want, speed does not matter.
You are not allowed to take notes nor to take pictures of the screen, etc. Please do not cheat!
For your information, you are not rewarded on your performance in this part, so there is no point in cheating.
First, you will go through a practice round in order to get used to the interface.
⟨Start practice⟩

Screen 5
[Practice on the memory ability task, see Figure B.1]

Screen 6
Part 1
You will now move to two main learning rounds (with no break in between).
There will be more pictures than in the practice round. The order will be the same in both learning rounds.
Just try to do your best so that we can assess as well as possible your learning capabilities.
Recall that you are not rewarded on your performance in this part.
⟨Start⟩

Screen 7
[Memory ability task (two rounds), see Figure B.2]

Screen 8
Part 1
We now ask you to report if you cheated in this part of the experiment (by taking a picture, taking notes, etc.).
Of course, using strategies such as telling yourself a story or remembering the first letters does not count as cheating.
If you did cheat, you will still be able to complete the rest of the experiment and earn the base payment as well as the bonus rewards. Your submission will not be rejected. However, it is very important for our research to know if you cheated in this part.

Please answer honestly: did you cheat on one of the learning rounds?
⟨No, I did not cheat⟩ ⟨Yes, but only on the last round⟩ ⟨Yes, in both rounds⟩
⟨Next⟩

Screen 9
Part 1: Reward
Congratulations, you completed Part 1!
As a reward for completion, you earn $4 as bonus payment. ⟨Next⟩

Screen 10
Part 2: Instructions
For Part 2, you will be randomly paired with another participant from Prolific.
They were recruited for the same experiment as you were, and they earned the same bonus.
The other participant and you will play a short game. You will play with them only once.
Current bonuses for both players:
You: $4
The other: $4
⟨Previous⟩ ⟨Next⟩

Screen 11
Part 2: Instructions
The game is as follows: one of you will be Player 1, the other Player 2.
Player 1 will take part in a lottery. They risk losing $4 from their bonus payment.
The lottery works in the following way: a ball will be drawn completely at random from an urn containing red and blue balls. The urn contains [21/24/27] balls in total.
If the ball drawn is blue, Player 1 will lose $4. If it is red, they won't lose anything.
Player 2 gets to decide how many red and blue balls there will be in the urn. They have two options:

- Option 1: 7 red balls, 14 blue balls
  - Player 1 is expected to lose around $2.67 on average
- Option 2: 14 red balls, 7 blue balls
  - Player 1 is expected to lose around $1.33 on average

Additionally, choosing Option 1 earns Player 2 an immediate bonus of $1.
Player 2's choice will be displayed to Player 1 before the ball is drawn.
⟨Previous⟩ ⟨Start⟩

Screen 12

Part 2

⟨Reveal my role⟩

Screen 13

Part 2

You have been randomly assigned to the role of Player 2
You get to decide which urn the other participant gets.
Choose carefully, as this game won't be repeated.
⟨Next⟩

Screen 14

[Dictator game variant, see Figure B.3]

Screen 15

Part 3: Questionnaire

In Part 3, you will be asked a few questions about yourself.
Please report your gender:
⟨Man⟩ ⟨Woman⟩ ⟨Other⟩ ⟨Prefer not to say⟩
Please report your age: [Number entry]
How do you see yourself: are you generally a person who is fully prepared to take risks, or do you try to avoid taking risks?
Not at all willing to take risks ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very willing to take risks
⟨Next⟩

Screen 16

Part 3: Questionnaire

Please indicate to what extent your participation in this experiment made you feel the following emotions:
I have felt happiness:
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
I have felt guilt:
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
I have felt pride:
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
I have felt annoyance:
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
⟨Next⟩

Screen 17

Congratulations!

You have successfully completed the first session.

The second session will become available on Prolific on January 29, 2026 at 3:00 PM PDT / 6:00 PM EDT.

It will be open for 8 hours. Make sure not to miss it!

As mentioned previously, you will receive your bonus payments upon completion of the second session.

You can leave fullscreen mode by pressing Esc.

Do you have any feedback that you want to give us on this study? (Things that were unclear, felt unfair or unrealistic, etc.)

[Free text box:]

⟨Complete study⟩

Follow-up session

Screen 18

Welcome back

Thank you very much for participating in the second session of this experiment!
You will receive your bonus payment from the first session upon completion.
You can also earn additional bonus payments in this session.
⟨Next⟩

Screen 19

Part 1

In Part 1 of this second session, you will be asked two questions on what happened in the first session.
For one question of these two questions (selected at random), you will receive a $0.5 bonus payment if your answer is considered valid (as defined on the question's page).
⟨Start⟩

Screen 20

Part 1

In the first session, you played a game with another participant.
The other participant took part in a lottery, from which they risked losing part of their bonus payment.
A ball was drawn completely at random from an urn containing red and blue balls. The urn contained 21 balls in total.
If the ball drawn was blue, the other participant lost $4 from their bonus payment. If it was red, they didn't lose anything.
You made a choice between two possible urns with different numbers of red and blue balls:

- Option 1: X red balls, Y blue balls
- Option 2: Y red balls, X blue balls

You earned an additional $1 for choosing the urn with more blue balls.
Please recall: what color was the ball drawn from the urn?
⟨blue⟩ ⟨red⟩

Screen 21

Part 1

In the first session, you played a game with another participant.
The other participant took part in a lottery, from which they risked losing part of their bonus payment.

A ball was drawn completely at random from an urn containing red and blue balls. The urn contained 21 balls in total.
If the ball drawn was blue, the other participant lost $4 from their bonus payment. If it was red, they didn't lose anything.
You made a choice between two possible urns with different numbers of red and blue balls:

- Option 1: X red balls, Y blue balls
- Option 2: Y red balls, X blue balls

You earned an additional $1 for choosing the urn with more blue balls.
Please recall: what color was the ball drawn from the urn?
⟨blue⟩ ⟨red⟩
Use the slider below to report how certain you are that the ball drawn was [selected color]:
50% [Slider] 100%
I am 50% certain that the ball was blue.
I am 50% certain that the ball was red.
If this question is the one rewarded, you will be rewarded according to a rule which makes it in your best interest to report what you think are the real chances that you are correct.
⟨More info⟩

Screen 22
[When clicking ⟨More info⟩ on the previous screen]
Reward mechanism
You will have to report a number P between 0 and 100, meaning that you believe that your chances of being correct are "P out of 100".
The computer will create a Lottery Bag:
The computer randomly chooses a number between 0 and 100. Each number is equally likely to be chosen. Although the computer knows this number, you do not. We call this randomly chosen number "?". The computer fills a bag with 100 chips. "?" chips are black and the rest are white. "?" in 100 chips are black.
There will be two ways to win a prize of $0.5: The "P" lottery, and the "Bag" lottery.
"P" lottery: win $0.5 if you are correct, and $0 if you are not. "Bag" lottery: A chip is drawn from the Lottery Bag. You win $0.5 if a black chip is drawn, and $0 if a white chip is drawn.
The computer knows your chance of winning $0.5 in the "Bag" lottery. Based on your reported belief P, the computer will select the lottery that gives you the highest chance of winning $0.5. (If the lotteries give you an equal chance of winning you will play the "Bag" lottery.)
You should think carefully about how likely you think it is that you are correct, as we will use your reported belief to decide whether you are paid according to one lottery or the

other.
Summary: There are two ways to win $0.5:
The "P" lottery awards $0.5 if you are correct, and $0 otherwise. The "Bag" lottery awards $0.5 if a black chip is drawn from the Lottery Bag. There is an unknown, random chance of winning $0.5.
Based on your reported belief P, you will play whichever lottery gives you the highest chance of winning $0.5.
This mechanism is designed so that it is in your best interest to report as your belief P what you think are the real chances that you are correct.
⟨Close⟩

Screen 23
Part 1
In the first session, you played a game with another participant.
The other participant took part in a lottery, from which they risked losing part of their bonus payment.
A ball was drawn completely at random from an urn containing red and blue balls. The urn contained 21 balls in total.
If the ball drawn was blue, the other participant lost $4 from their bonus payment. If it was red, they didn't lose anything.
You made a choice between two possible urns with different numbers of red and blue balls:

- Option 1: X red balls, Y blue balls
- Option 2: Y red balls, X blue balls

You earned an additional $1 for choosing the urn with more blue balls.
Please recall: in the urn which you chose, how many balls were blue?
There were [number field] blue balls out of 21 in the urn (and [21 - number field] red balls)
For this question, only exact answers are eligible for the bonus payment.
⟨Confirm⟩

Screen 24
Part 1
In the first session, you played a game with another participant.
The other participant took part in a lottery, from which they risked losing part of their bonus payment.
A ball was drawn completely at random from an urn containing red and blue balls. The urn contained 21 balls in total.
If the ball drawn was blue, the other participant lost $4 from their bonus payment. If it was red, they didn't lose anything.

You made a choice between two possible urns with different numbers of red and blue balls:

- Option 1: X red balls, Y blue balls
- Option 2: Y red balls, X blue balls

You earned an additional $1 for choosing the urn with more blue balls.
Please recall: in the urn which you chose, how many balls were blue?
There were [number field] blue balls out of 21 in the urn (and [21 - number field] red balls)
For this question, only exact answers are eligible for the bonus payment.
⟨Confirm⟩

Screen 25
Part 1
In the first session, you played a game with another participant.
The other participant took part in a lottery, from which they risked losing part of their bonus payment.
A ball was drawn completely at random from an urn containing red and blue balls. The urn contained 21 balls in total.
If the ball drawn was blue, the other participant lost $4 from their bonus payment. If it was red, they didn't lose anything.
You made a choice between two possible urns with different numbers of red and blue balls:

- Option 1: X red balls, Y blue balls
- Option 2: Y red balls, X blue balls

You earned an additional $1 for choosing the urn with more blue balls.
Please recall: in the urn which you chose, how many balls were blue?
There were [number field] blue balls out of 21 in the urn (and [21 - number field] red balls)
For this question, only exact answers are eligible for the bonus payment.
How confident are you that you correctly remember the urn's content?
Not at all confident ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ Very confident
⟨Confirm⟩

Screen 26
Part 2
In the first session, you had to learn the order of a list of pictures.
Now, you will have to try to remember which objects were present or not in this list.
You will see pictures presented one by one, and you will have to answer in limited time whether you saw this object during the first session.
To answer quickly, you will use your keyboard. You will have to press the Space key when a picture is presented to report that it was in the list from the first session.

There is no bonus payment for this part.
⟨Start demo⟩

Screen 27
[Demo on follow-up of memory ability task, see Figure B.4]

Screen 28
Part 2
You will now start the actual task.
Select the object with Space if you saw it during the first session
⟨Start⟩

Screen 29
[Follow-up of memory ability task, see Figure B.5]

Screen 30
Part 3
In this part, you will have to allocate up to $15 between two other participants.
The two participants will be:
The participant who played with you as Player 1 in the lottery game in the previous session. Another participant who played as Player 1 in the lottery game in the previous session (randomly selected).
You can allocate the money as you want between the two participants, as long as the sum doesn't go beyond the maximum of $15.
This decision involves high potential bonus payments, but the bonus payments will not be implemented for everyone in the study. Only one randomly selected pair of participants will receive the bonus payments related to this part.
Thus, your decision has a chance to be really implemented. In that case, any amount not allocated to the other two participants will be kept by the researcher.
⟨Start⟩

Screen 31
Part 3
Indicate your preferred allocation by typing the amounts or using the sliders:
Other Player 1 (randomly selected) [Input slider from $0 to $15]
Player 1 with whom you played [Input slider from $0 to $15]
Total allocated [Output slider from $0 to $15, displaying the sum of the other sliders]
[Checkbox] I understand that this allocation has a chance to really be implemented.
⟨Confirm⟩

Screen 32

Part 4 - Instructions

In this part, you will play a game with another participant from Prolific (with whom you have not yet interacted).

The game involves high potential bonus payments, but the bonus payments will not be implemented for everyone in the study. Only one randomly selected pair of participants will receive the bonus payments related to this game.

You will have no further interaction with the other participant after the game.

⟨Next⟩

Screen 33

Part 4 - Instructions

The game is as follows:

The other participant will take part in a lottery where they might earn $12.

A ball will be drawn completely at random from an urn containing orange and green balls. The urn contains 3 balls in total.

If the ball drawn is green, they will earn a bonus payment of $12. If it is orange, they won't earn anything.

You get to decide how many green and orange balls there will be in the urn. You have two options:

- Option 1: 2 green balls, 1 orange ball
  - The other participant is expected to earn around $8.0 on average
- Option 2: 1 green ball, 2 orange balls
  - The other participant is expected to earn around $4.0 on average

You can also earn a bonus payment for yourself from both options, as described in the next screen.

⟨Previous⟩ ⟨Next⟩

Screen 34

Part 4 - Instructions

Option 1 earns a fixed bonus payment of $8.

Option 2 earns a bonus payment which is independent of the lottery, but can take different values. In the rest of the instructions, it is symbolized by a money bag [money bag].

You will have to report whether you would like to choose Option 1 or Option 2 depending on the value of [money bag].

To do so, you will report your choice between the options for each possible value in a table like the one below: You earn... Option 1 $8 $8 $8 $8 $8 $8 Option 2 You earn... [money

bag] = \$8 [money bag] = \$9 [money bag] = \$10 [money bag] = \$11 [money bag] = \$12 [money bag] = \$13
In the end, if your pair is the one randomly selected to receive the bonus payments corresponding to this game, the value of [money bag] will be one of those possible values (selected at random). A ball will be drawn from the urn corresponding to your choice for that value.
⟨Previous⟩ ⟨Start⟩

Screen 35
[Multiple-price list version of the dictator game variant, see Figure B.6]

Screen 36
Part 5
From what you can recall, please report to what extent your participation in the first session made you feel the following emotions:
In the first session, I have felt happiness:
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
In the first session, I have felt guilt:
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
In the first session, I have felt pride:
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
In the first session, I have felt annoyance:
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
⟨Next⟩

Screen 37
Part 5
Have you thought about the first session since you completed it?
⟨Not at all⟩ ⟨Once or twice⟩ ⟨Several times⟩ ⟨Quite a lot⟩
How much time did you spend on Prolific since the last session? (on any device):
⟨Less than 1 hour⟩ ⟨1-2 hours⟩ ⟨2-5 hours⟩ ⟨More than 5 hours⟩
⟨Next⟩

Screen 38
Part 5
How much did you enjoy taking part in this experiment overall?
Not at all ⟨0⟩ ⟨1⟩ ⟨2⟩ ⟨3⟩ ⟨4⟩ ⟨5⟩ ⟨6⟩ ⟨7⟩ ⟨8⟩ ⟨9⟩ ⟨10⟩ Very much
Have you done anything special which helped you remember the first session? (Taken notes, etc.) Your participation will be validated and your bonus paid independently of your answer, so you can answer fully honestly.

[Free text box:]

Screen 39
The End

Thank you very much for your participation, you helped science today!
You will receive your bonus quickly once all participants are done with the session.
Do you have any feedback that you want to give us on this experiment? (Things that were unclear, felt unfair or unrealistic, etc.)
[Free text box:]
⟨Return to Prolific⟩