2-28-2022

Virtue Preferences: Jekyll and Hyde Paradoxes with Sanctions

James Konow
Loyola Marymount University, jkonow@lmu.edu

FRANK COMMENTS WELCOME (NOW BEFORE YOU REFEREE IT)

February 2022

Virtue Preferences: Jekyll and Hyde Paradoxes with Sanctions

James Konow
Loyola Marymount University
One LMU Drive, Suite 4200
Los Angeles, CA 90045-2659
Telephone: (310) 338-7486
Email: jkonow@lmu.edu
Website: jameskonow.com

Abstract

Jekyll and Hyde paradoxes refer to the fact that people sometimes behave morally in certain situations but then behave immorally (or, at least, less morally) under conditions that differ for reasons that seem morally irrelevant. Observational and experimental studies confirm the economic and social importance of these phenomena, which are inconsistent both with rational self-interest as well as with theories that add stable moral preferences. This paper presents a theory that reconciles various of these phenomena, including the depressing effects on moral behavior of experimentally introducing options to take the earnings of others, to delegate decisions and to remain ignorant of the consequences of one's decision, as well as rewarding and punishing others for uncontrollable luck. The theory introduces the concept of virtue preferences, which together with a model combining moral salience, fairness and altruism, explain not only these paradoxes but also classic findings on reciprocity. The results of an experiment that tests the theory out-of-sample prove consistent with the theoretical predictions.

Keywords: moral salience, virtue preferences, conditional altruism, fairness, altruism, reciprocity, moral wiggle room

JEL Classification: C9, D3, D9

Acknowledgements: I wish to thank Joel Sobel for comments and suggestions, Prachi Jain for advice, and Veronica Backer-Peral, Ben Mouehli, Gerrit Nanninga, and Eamon Shaw for their research assistance.

1. Introduction

Companies hire consulting firms to recommend or carry out the firing of their employees, although the companies could implement the firings themselves and save the consulting fees. In the lead-up to the 2007-08 financial crisis, lenders, who were formerly prudent, chose to avoid documenting applicants' incomes. In the early 2000s, CEOs of troubled firms like Enron and Worldcom professed ignorance of the dubious accounting practices at their companies. An extensive literature in economics on reciprocity shows people reward good behavior and punish bad behavior, but more recent research finds that delegating decisions or remaining willfully ignorant, as in the above examples, sometimes enables decision-makers to avoid accountability. Moreover, even when decision-makers are held accountable, other research shows that such reckoning is not always limited to actual choices: people reward or punish politicians and CEOs not only for their good or bad choices but sometimes also for uncontrollable luck.

Various explanations can be offered for these "anomalies," such as social image concerns, specialization, changes in regulatory regimes, fear of prosecution, risk preferences, or other (strategically) self-interested considerations. But the results of more recent laboratory and field experiments that constitute experimental renditions of these paradoxes demonstrate their robustness to careful controls. Diverse theories have been designed to explain many of the "classic" findings from decades of research based on a combination of self-interest and stable social preferences. The more recent anomalies, however, point to inconsistencies, not only with unadulterated self-interest, but also with the prevailing social preference theories. For instance, self-interested individuals should take the selfish actions, e.g., firing employees themselves or revealing information that identifies the most self-serving choice, and morally motivated decision-makers should take the moral actions, e.g., not firing the employees or revealing information that identifies the most moral action, but neither selfish nor moral types should delegate decisions or avoid information. Moreover, social preference theories that incorporate reciprocal motives to sanction, i.e., to reward or punish, do not predict or explain the effects cited above on sanctioning behavior involving delegation, willful ignorance or uncontrollable luck.

This paper presents a theory of moral salience, conditional altruism and virtue preferences that accounts for various classic as well as more recent anomalous findings on moral preferences. It introduces a theory of sanctioning called virtue preferences and integrates it into the theory of allocative preferences introduced in the paper, "Moral Salience and Conditional

Altruism: Reconciling Jekyll and Hyde Paradoxes" (henceforth Konow, 2022). It applies the theory to additional stylized facts of experiments not covered in the other paper and reports the results of an original experiment that tests and finds support for the theory out-of-sample. The theory is related to the oldest school of thought in Western moral philosophy, virtue ethics. With reference to this school, Ashraf and Bandiera (2017) explore how altruistic acts affect altruistic capital, and Konow and Earley (2008) discuss the relationship between virtue and happiness. The current theory relates to other features of virtue ethics, including its claims of moral pluralism (i.e., the existence of multiple moral principles), context-dependent morality, and a concern for virtues.

The theory is formulated here to address certain types of decisions. Although strategic decision-making is obviously economically important (and reference will sometimes be made to results on such decisions), formally I focus on the unilateral choices of a decision-maker, or *agent*, acting on a passive person, or *patient*. There are several reasons for this focus: it helps zero in on moral motives while avoiding the potentially confounding effects of strategic self-interest, it facilitates the parallel development of a simple model and experimental design to test the model, and it relates to experimental and observational evidence on such decisions from a large (although not unwieldy) literature. Anonymous donations to charities provide an example from the field of the kind of decisions considered. An example from the laboratory is the dictator game: in one variation on this design, two subjects are endowed with fixed sums of money, and the subject with a larger endowment, the *dictator*, may anonymously transfer amounts to the subject with a smaller endowment, the *recipient*, who has no recourse. An anomaly discussed in Konow (2022) and later in this paper is the "taking effect:" in a standard dictator game, the dictator may only transfer a non-negative amount to the recipient, but if options to take from the recipient are added to an otherwise equivalent treatment, there is a reduction in the fraction of dictators who give and in the average transfer among those who give. This taking effect is inconsistent with stable moral preferences: a dictator, who is selfless enough to give in the standard case, should give the same amount, if taking is permitted.

Consider this rough sketch of the theory, beginning with the framework introduced in Konow (2022). In addition to material utility, the agent is assumed to have allocative preferences, specifically, called conditional altruism, which consists of fairness preferences and altruism. Fairness preferences represent the disutility to the agent of the patient receiving more or

less than the patient's fair payoff. Altruism is (in the case of most agents) the utility from transferring an amount to the patient, or, where relevant, the agent's disutility of taking from the patient. Note that, although suitable for the contexts considered here, these simple preferences are not presumed to exhaust all moral preferences. Allocative preferences are weighted by moral salience, which represents the prominence of moral preferences in the agent's utility function. Moral salience is a function of the decision context, i.e., of the agent's set of choices and information about the choices. This weight increases with moral context, e.g., if the agent is permitted to share more money, and decreases with non-moral context, e.g., if the agent is permitted to take more money from the patient. The obvious implication of these assumptions is consistent with the taking effect described above: adding taking options reduces the prominence of, and weight on, moral preferences and, therefore, dictator giving.

This paper extends this framework by introducing virtue preferences, which represent the desire to reward or punish another beyond what is called for by fairness alone. For concreteness, think again of the dictator game described above but with an additional second, unannounced stage involving an anonymous allocator. In contrast to the first stage in which the dictator is a *stakeholder*, whose own earnings are affected by his or her decision, in this second stage the allocator is a third-party *spectator*, who receives a fixed payment to distribute an additional sum of money between the dictator and the recipient from the first stage. Virtue preferences are incorporated as an argument in fairness preferences, in this case, those of the spectator, which adjusts the payoff to the dictator that minimizes inequity aversion for the spectator, upwards in the case of rewarding the dictator and downwards in the case of punishing the dictator. The ideal adjustment is a function of both the action and intrinsic motivation of the dictator. That is, in this application, the agent's desire to sanction is not a function of the dictator's intrinsic motivation alone, if that motivation is not accompanied by action on the part of the dictator. So, a generous or selfish dictator is not rewarded or punished, respectively, merely for his or her moral preferences. It is also not a function of the dictator's action alone. To the latter point, consider a given dictator transfer, say of $8, in the standard game without taking options. That same dictator is predicted to transfer less, if taking options are added, because of reduced moral salience. Thus, a dictator who transfers $8 in the treatment with taking signals greater intrinsic generosity than the one who transfers $8 in the standard treatment and is predicted, therefore, to be treated more favorably by a spectator. The ideal sanction, therefore, is a function of both the action and what it

conveys in the given context about the strength of the dictator's intrinsic moral motivation.

Section 2 presents the theory with especial attention to virtue preferences, which is then applied to classic results on reciprocity in section 3 and to the anomaly of outcome bias in section 4. Section 5 applies the theory to an experiment similar to the scenario just described that tests the theory out-of-sample and relates the results to the taking effect and to variations in sanctioning with taking options. The next two sections address two types of norm avoidance, viz., willful ignorance in section 6 and delegation in section 7, and propose explanations for stakeholder allocations and norm avoidance and for third party sanctions. Section 8 discusses briefly a different type of moral salience called point salience, and section 9 concludes.

2. Theory

This section presents the general theory of moral salience, conditional altruism and virtue preferences. As moral salience and conditional altruism are described in detail in Konow (2022), I refer the reader to that paper for a more in-depth treatment and discuss them more succinctly here, focusing attention on the introduction of virtue preferences and their integration into the general model.

The agent may choose an action, *x*, from the set of available actions, X, e.g., such as dictator's transfer to a recipient or a donor's gift to a charity. The agent's decision might also be impacted by other potentially morally relevant information, e.g., knowledge of a recipient's endowment, *y*, as well as other elements of the set of information related to the decision, Y. Together, X and Y form the decision context, C, which may be partitioned in a different manner into moral context, C_+, and non-moral context, C_-. There are measures on these subsets, $m(C_i)$, whereby the moral measure is denoted $p = m(C_+)$, and the non-moral measure is denoted $n = m(C_-)$. Moral salience, $\sigma(p, n)$, is the weight applied to the agent's moral preferences and is a function of the measures of moral and non-moral context, viz., σ is increasing in p and decreasing in n. In general, moral salience is decreasing in factors that increase the perceived separation between the agent's choice and the moral consequences of that choice on the patient. In the taking game, for example, expanding opportunities to give increases moral context, C_+, its measure, p, and, therefore, moral salience, i.e., it increases the weight on moral preferences. Conversely, expanding opportunities to take increases non-moral context, C_-, increases its measure, n, and, therefore, decreases moral salience. Moral salience is a type of *set* salience,

which is a function of measures on subsets and captures the tendency for the subset with smaller measure to have disproportionate prominence relative to the contrasting subset with larger measure. This is distinct from moral *point* salience, which relates to elements of sets and is discussed in section 8.

Moral salience maps the moral and non-moral measures of the decision context into the half-open unit interval, $\sigma: \mathbb{R}_+^2 \to (0,1]$, whereby I assume that $\sigma(p,0) > 0, p > 0;\ \sigma(0,n) \geq 0, n > 0$; and that $\sigma(p,n)$ is twice continuously differentiable with

$$\frac{\partial \sigma}{\partial p}\Big|_{n>0} > 0, \frac{\partial^2 \sigma}{\partial p^2}\Big|_{n>0} < 0, \frac{\partial \sigma}{\partial n}\Big|_{p>0} < 0, \frac{\partial^2 \sigma}{\partial n^2}\Big|_{p>0} > 0.$$

A convenient specification that captures the assumed relationships is the ratio $\frac{p}{p+n}$. Further, many contexts contain baseline moral salience, i.e., fixed moral considerations activated by being in certain situations, such as being a dictator in a dictator game. This baseline salience is denoted $\bar{\sigma} \in [0,1)$ and leads to the following specification for moral salience:

$$\sigma = (1 - \bar{\sigma}) \cdot \frac{p}{p+n} + \bar{\sigma}.$$

For this ratio, I additionally assume $p > 0$ guaranteeing that $\sigma > 0$. The proposed conditions capture the tendency for the subset with smaller measure to have disproportionate prominence as well as a further property: the marginal salience of the first addition of moral context is greater than that of the second, i.e., moral salience increases at a decreasing rate; conversely, the effect of the first addition of non-moral context is greater than the second, i.e., moral salience decreases at a decreasing rate. For example, allowing a dictator to take \$2 from a recipient reduces moral salience more than allowing the dictator to take only \$1, but the reduction in salience occasioned by the first dollar of taking is greater than that of the second.

Moral salience is a weight applied to moral preferences, here using the model of *conditional altruism*. This model has three components: material utility and two moral motives, fairness and altruism. Material utility, $u(\pi_a)$, is assumed to be a twice continuously differentiable function of the agent's material allocation, π_a, whereby $u(0) = 0, {}^{\partial u}/_{\partial \pi_a} > 0$, and ${}^{\partial^2 u}/_{\partial {\pi_a}^2} \leq 0$. Fairness, $f(\pi_p - \eta)$, captures the disutility experienced by the agent, as the patient's allocation, π_p, differs from the fair allocation to the patient or so-called *entitlement*, η. This term is strictly concave with a maximum where $\pi_p = \eta$, viz., $f(0) = 0$, ${}^{\partial f}/_{\partial w} \cdot w < 0$ for

$w \equiv \pi_p - \eta \neq 0$, and $\partial^2 f / \partial w^2 < 0$. A fairness coefficient, $\phi > 0$, which captures differences across agents in the strength of their fairness preferences and is applied to f to form $\phi f(\pi_p - \eta)$. Altruism is a twice continuously differentiable function, $g(x, \alpha)$, of the amount given or taken by the agent, *x*, and an altruism coefficient, α, where $g(0, \alpha) = 0$, $\partial g / \partial x \gtreqless 0$ as $\alpha \gtreqless 0$, $\partial^2 g / \partial x^2 < 0$, $\partial g / \partial \alpha \cdot x > 0$ for $x \neq 0$, and $\partial^2 g / \partial x \partial \alpha > 0$. Specifically, agents differ according to the value of their altruism coefficient, α, and can be categorized as altruistic, $\alpha > 0$, selfish, $\alpha = 0$, or spiteful, $\alpha < 0$, and the slope of g is increasing in α. Note that this resembles warm glow, when $\alpha > 0$ and $x > 0$ (e.g., Andreoni, 1989), but it also encompasses self-interest, spite and taking (i.e., $x < 0$). Altruism, in this model, is personal and unlike pure altruism in two senses. First, it is a function solely of that part of the patient's allocation that can be attributed to a personal choice of the agent, e.g., a dictator transfer to or from a recipient, and not of the patient's total allocation. Second, it is personal in the sense of applying to partial relationships, i.e., to agent-patient relationships but not to impartial third-party, or spectator, decisions.

Assuming additively separable utility and weighting fairness and altruism by moral salience, the utility of the agent, U, is

(1) $\quad U = u(\pi_a) + \sigma\phi f(\pi_p - \eta) + \sigma g(x, \alpha).$

We will examine several variations on the dictator game described in the introduction, so note that (1) may, in this case, be written

(2) $\quad U = u(X - x) + \sigma\phi f(Y + x - \eta) + \sigma g(x, \alpha),$

where *X* is the endowment of the dictator, *x* the dictator's transfer to the recipient such that $\pi_a = X - x$, and *Y* the recipient's endowment such that $\pi_p = Y + x$. Below I mention only briefly the most important additional assumptions about this model and refer the reader to Konow (2022) for more detailed discussion.

Context enters into the theory in a second way in addition to moral salience, namely, through the entitlement. A substantial experimental literature demonstrates that distributive goals can correspond to different norms or combinations of norms depending on the decision context. For instance, efficiency is relevant when total surplus can vary, equity (i.e., proportionality) contributions chosen by individuals differ, need where there is information about basic needs, and equality by default when morally relevant information about other norms is lacking (see

Konow, 2003, and Konow, Saijo and Akai, 2020). I discuss plausible justifications for different entitlements in the various experiments treated based on lessons from these stakeholder and spectator decisions. The latter, in particular, provide information on norms that, compared to the former, is undistorted by self-interested bias and by noise caused by individual differences in the degree of self-interest (see Konow, 2005, 2009). In cases involving simple dictator decisions over fixed sums, the entitlement is assumed fixed. Although theoretical claims do not generally depend on this, I also assume it reduces to equal splits in such games, both for concreteness and for consistency with evidence on spectator decisions from such experiments.

When uncertainty is involved, I assume agents are expected utility maximizers. I also assume that $Cov(\alpha, \phi) > 0$, i.e., fairer agents are, on average, more altruistic. Although not necessary to explain most allocative behavior, it is a reasonable assumption that later proves useful for certain claims involving virtue preferences. Konow (2022) describes in detail other assumptions that ensure the model is consistent with intuition and stylized results from many experiments, including that, in the standard dictator game, some dictators act selfishly, and that minimally fair, spiteful, and super-fair dictators (i.e., ones who transfer more than a fair share) are all minorities.[1] Here I summarize briefly some claims, which are stated and discussed in greater detail in that paper, identifying them by their theorem numbers there. The optimal transfer is increasing in σ, α, η, and ϕ, except for being decreasing in ϕ with super-fair agents (2.1, 2.4, 2.5, and 2.3, respectively). The optimal transfer is decreasing in the measure of non-moral context, n, and, assuming the optimal transfer is weakly convex in σ, is strictly convex in n (2.2). This last claim is consistent with empirical results presented in Konow (2022) and is important for the focus of that paper and this paper on variation in n.

The utility function presented thus far is, in philosophical terms, entirely consequentialist, i.e., it represents preferences over outcomes or the consequences of decisions. Material utility reflects the material allocation of the agent. Conditional altruism, captured by fairness and altruism, are allocative preferences with respect to the fair transfer and the agent's endowment,

[1] Formally, the assumptions are as follows: ϕ is distributed according to the cdf $\Phi(\phi)$, which has support $[\underline{\phi}, \overline{\phi}]$ with $0 < \underline{\phi} < \overline{\phi} < \infty$ and $0 < \Phi\left(\underline{\phi}\right) < 0.5$, α is distributed according to the cdf $\mathrm{A}(\alpha)$, which has support $[\underline{\alpha}, \overline{\alpha}]$ with $-\infty < \underline{\alpha} < 0 < \overline{\alpha} < \infty$, $\mathrm{A}(0) < 0.5 < \mathrm{A}(\overline{\alpha}) - \mathrm{A}(0)$, $\int_{\underline{\alpha}}^{\overline{\alpha}} \alpha\, \rho(\alpha) d\alpha > 0$ where $\rho(\alpha)$ is the probability density function of α, and $0 < \alpha^* < \overline{\alpha}$ and $0 < \mathrm{A}(\overline{\alpha}) - \mathrm{A}(\alpha^*) < 0.5$ where σ^* is the level of salience in the standard dictator game and $\alpha^* = \{\alpha | u'(X - \eta) = \sigma^* \alpha g'(\alpha \eta)\}$. Finally, for even the most spiteful dictator (with $\underline{\alpha}$), $\underline{\alpha} g'(\alpha x^*) > -\phi f'(\phi(Y + x^* - \eta))$, where x^* is the optimal transfer in the standard dictator game.

respectively. Now I introduce virtue preferences, which represent an additional non-consequentialist moral motive, viz., to *sanction*, that is, to reward or punish others.[2]

Most theoretical accounts of sanctioning are *reciprocity* theories, e.g., Charness and Rabin (2002), Falk and Fischbacher (2006), Rabin (1993), and Dufwenberg and Kirchsteiger (2004). These theories formulate a motive to reward or punish others based on their so-called intentions. Good or bad intentions are inferred based on both others' (expected) choices and their available choice set. For example, the standard formulation considers whether the expected consequences of another's action exceed or fall short of some "fair" benchmark, whereby the benchmark is defined relative to the other's available choice set. An alternative approach is to formulate this motive with respect to *moral types*. Levine (1998) introduced such a model in which the sanctioning motive depends jointly on the altruism (or spite) of both the agent and the patient. In his model, an agent might be altruistic or spiteful but is more altruistic (spiteful) toward a more altruistic (spiteful) patient.

Virtue preferences incorporate core feature of virtue ethics, which is the oldest school in Western moral philosophy. Its advocates span millennia and include Aristotle (1925), Adam Smith (1759), and Martha Nussbaum and Amartya Sen (1993). It should be noted that there are different schools of thought within virtue ethics, but the theoretical concepts presented here are variations on common positions that can be found in that branch of ethics. Morality is viewed as pluralistic, i.e., consisting of multiple principles and preferences, whereby in the present case of allocative preferences, these constitute fairness and altruism.[3] As interpreted in the present theory, a virtue is a willingness to act on moral principles for the benefit of others, e.g., to be fair or altruistic to others, that is actually acted on. This differs from the approaches reviewed above in subtle but important ways. Reciprocity theories depend solely on consequences or intended consequences, but not so with virtue, which depends on moral motivation. Virtue shares with the moral types approach its link to moral preferences, but it is also distinct in several ways. First, moral preferences must be realized in action, and, unlike the moral types approach, the relevant metric is behavioral and not the latent variable of moral preferences. Second, and in a related

[2] Although we do not explore this aspect here, in a dynamic framework, virtue preferences might serve to undergird virtue itself, indeed, Dal Bó and Dal Bó (2014) find that punishment is a critical force that sustains the favorable effects of increased moral salience on cooperation.

[3] In the standard Aristotelian terminology, these two virtues correspond to justice and liberality, respectively, although Aristotle also discussed other virtues, among them prudence, courage, truthfulness, and friendship.

point, the sanctioning motive in virtue preferences is not over the moral preferences of others but over their actions that benefit (or harm) others. For example, in a dictator game, fair dictators are not rewarded for their intrinsic altruism per se but for generous behavior that manifests their moral preferences. Neither intent alone nor action alone suffices: rather, virtue is intent (understood now as moral preferences) coupled with action. Finally, a subtle but important point is that virtue is the *notional* willingness to behave morally, even if the *effective*, or actual, behavior differs due to obstacles that preclude more precise expression of that willingness because choice is subject to constraints or uncertainty. For example, a dictator, who is willing to share $12 with a recipient is more virtuous than a dictator, who is only willing to share $10, even if both share the same $10 due to experimental rules that cap transfers at $10. Note, however, that, to count as virtue, action must be involved, even if the effective action differs from the notional one.

Moral character refers to an individual's set of virtues. Of course, depending on the context, multiple virtues might be in play. For example, in the contexts examined here, conditional altruism predicts that moral behavior reflects equity and altruism, and later examples add efficiency. Virtue ethicists argue that the relative importance of the different virtues in determining right action is context-dependent. The theory presented here also involves context-dependence: the context determines which moral norms are relevant in identifying the entitlement, and moral and non-moral context determine the relative and absolute salience of morality as a whole. Note that virtues are valued only in so far they benefit others, so in the context of allocative preferences, I assume moral character can be measured by intrinsically motivated *generosity*, that is, the willingness to sacrifice materially, whether for the sake of fairness, altruism or both. Denote this γ, where $\gamma \in \mathbb{R}$, and suppose γ is distributed according to the cumulative distribution function $\Gamma(\gamma)$, where $\Gamma(\gamma)$ has support $[\underline{\gamma}, \overline{\gamma}]$ with $-\infty < \underline{\gamma} \leq 0 < \overline{\gamma} < \infty$. This is the variable that individuals are assumed to be motivated to sanction in the contexts studied here.[4] Specifically, the morally relevant generosity is notional, so I will conceptualize it with the following *reference state*. Consider an agent, who may make a unilateral, anonymous and unlimited transfer of resources to or from a patient. There is no role

[4] Note that the moral types formulation produces a counterintuitive implication for fairness preferences with super-fair agents: as previously mentioned, in this case, a fairer agent is less generous but in the moral types approach more deserving of reward. Formulating the variable that is sanctioned in behavioral terms, in which only disadvantageous inequity aversion contributes to generosity, avoids this unfortunate implication.

for strategic self-interest, and the agent's only motivation for departing from narrow material interests is moral preferences. The reference state is like a standard dictator game, except that the dictator has unlimited access to the entire material endowments of both parties. Further assume that, in this hypothetical dictator game, even the most generous dictator would keep some his or her own endowment ($\overline{\gamma} < X$) and that even the most selfish dictator would not take all of the recipient's endowment ($\left|\underline{\gamma}\right| < Y$). Although not necessary, this assumption simplifies the analysis, but it also seems plausible, at least for populations like subjects in economics experiments, who would likely neither give away their last material possession nor take away the last possession of another.

Virtue preferences are preferences over moral character, specifically, preferences to reward the good, or praiseworthy, moral character of another, or to punish the bad, or blameworthy, moral character of another. In the present case, moral character is an expression of allocative preferences, which reduces to intrinsic generosity, so that virtue preferences are preferences over this generosity. These sanctioning preferences consist of several parts. Let us begin with an agent, who is capable of sanctioning others. For concreteness, think of a spectator, who may make transfers to or from a dictator after observing that dictator's transfer to a recipient. That is, the agent may transfer an amount to or from another, denoted $z \in \mathbb{R}$, whereby the range of possibilities in studies considered here comprises $z \in [\underline{z}, Z]$, where $-\infty < \underline{z} \leq 0$, $0 \leq Z < \infty$ and $\underline{z} < Z$. This agent may make transfers for allocative reasons but also in order to sanction, i.e., to reward or punish another beyond what allocative preferences alone demand. The agent's ideal level of sanctioning is denoted $\tilde{z}$, and this depends on the agent's estimate of the other's moral character (in the present case, the agent's estimate of the sanctioned person's notional generosity), $\hat{\gamma}$. Since the decision context differs from the thought experiment described above, the other's actual moral character, γ, is not known and must be estimated. As explained in later sections, $\hat{\gamma}$ is based on the other's choices as well as the decision context, including the reigning moral salience and constraints on the other's choices. In addition, for each agent and given the level of salience in the reference state, $\tilde{\sigma}$, there is a threshold value of γ, denoted $\tilde{\gamma}$, where I assume $\underline{\gamma} < \tilde{\gamma} < \overline{\gamma}$. Above this threshold, the agent judges the other's character as praiseworthy and deserving of reward and, below it, the other's character is viewed as blameworthy and deserving of punishment. This "character threshold" may differ across

sanctioning agents.

The ideal sanction, $\tilde{z}$, is assumed to depend on $\tilde{\gamma}$ and $\hat{\gamma}$ through the term $r\colon \mathbb{R}^2 \to \mathbb{R}$, which is the twice continuously differentiable function

$$r(\hat{\gamma} - \tilde{\gamma})$$

where $r(0) = 0, {}^{\partial r}/_{\partial k} > 0$ for $k \equiv \hat{\gamma} - \tilde{\gamma}$, and ${}^{\partial^2 r}/_{\partial k^2} < 0$.

This implies a positive ideal sanction, $\tilde{z} > 0$, when estimated character exceeds the threshold, $\hat{\gamma} > \tilde{\gamma}$, and a negative $\tilde{z} < 0$, when the opposite is the case, $\hat{\gamma} < \tilde{\gamma}$. In addition, the concavity of r captures the idea that blameworthy character implies greater punishment than the reward for praiseworthy character of an equal degree. I assume that agents care in differing degrees about sanctioning, denoted $\theta \geq 0$, which is distributed according to the cumulative distribution function $\Theta(\theta)$, where $\Theta(\theta)$ has support $[\underline{\theta}, \overline{\theta}]$ with $0 = \underline{\theta} < \overline{\theta} < \infty$, and $0 < \Theta(0) < 1$. That is, there is a mass of people, who care not a whit about sanctioning. To r and θ we add the scale of the sanction, x', which specifies the magnitude of reward or punishment appropriate to the context. This provides a measure of the importance of the action. To ignore the scale of the decision context, when choosing how to sanction character, would have implausible implications, such as taking a bad person, who is curt with a waiter, to be equally deserving of punishment as someone of equally bad character, who robs a bank. Since virtue is willingness coupled with action, and people are not sanctioned merely for being, character must be matched with a context that involves choice. In contexts with certainty about the mapping of choices to allocations, I propose defining the scale as the transfer called for by the moral norm, viz., the patient's entitlement less any endowment such that $x' = \eta - Y$ (I will present a more general specification that includes uncertainty in section 4). Thus, the ideal sanction can be expressed

$$\tilde{z} = \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot x'.$$

If a person is precluded from taking an action that reveals moral character, then I assume $r \equiv 0$ and, therefore, $\tilde{z} = 0$.

Finally, to sanction means to increase or decrease the patient's payoff beyond what is called for by distributive preferences alone (i.e., η), so $\tilde{z}$ is incorporated into the fairness function. The complete specification of the utility function of an agent with moral salience, material utility, fairness, altruism and virtue preferences is

$$U = u(\pi_a) + \sigma\phi f(z - \eta - \tilde{z}) + \sigma g(z, \alpha). \tag{3}$$

We next proceed to flesh out various specifications of this utility function to classic results on reciprocity and then apply it to a new experiment followed by three anomalies.

3. Reciprocity

Reciprocity refers to a type of behavior, where people may deviate from fairness in order to return kindness with kindness, called positive reciprocity, or unkindness with unkindness, called negative reciprocity. Such behavior has been found in numerous experimental designs, including with the seminal gift exchange game of Fehr, Kirchsteiger and Riedl (1993), the trust game of Berg, Dickhaut and McCabe (1995), the triadic design of Cox (2004), and the moonlighting game of Abbink, Irlenbusch and Renner (2000). These results can be summarized in the following stylized fact (abbreviated SF) from experiments on reciprocity.

SF 3.1: Stakeholders punish selfish and reward generous choices (e.g., Güth, Schmittberger and Schwarze, 1982, Berg, Dickhaut and McCabe, 1995, Cox, 2004, Abbink, Irlenbusch and Renner, 2000). Moreover, they sanction asymmetrically, punishing selfishness more strongly than they reward an equal degree of generosity (e.g., Croson and Konow, 2009, Cushman, Dreber, Wang and Costa, 2009, Offerman, 2002).

Further studies show that subjects exhibit generalized reciprocity, acting not only when they are themselves the objects of kindness or unkindness but also as third parties sanctioning kindness or unkindness by others toward others, e.g., Almenberg, Dreber, Apicella and Rand (2011), and Fehr and Fischbacher (2004).

In strategic designs, it is often ambiguous, whether such behavior reflects reciprocal altruism, i.e., a preference to reward or punish, or some other motive(s) (see Sobel, 2005, for an excellent theoretical treatment of types of reciprocity).[5] For that reason and others, I frame the

[5] As an illustration of this point, consider the ultimatum game, in which a "proposer" proposes a division of a fixed sum of money between him/herself and a "responder," and the responder either accepts, and the sum is divided as proposed, or rejects, in which case both earn nothing (Güth, Schmittberger and Schwarze, 1982). The subgame perfect Nash equilibrium under the standard assumptions of rational, self-interested agents is for the responder to accept any amount, however small, and for the proposer, therefore, to offer the minimum amount. Nevertheless, the results of hundreds of replications show that proposers typically offer non-negligible positive amounts and responders often reject positive offers of less than one-half (Camerer, 2003). But there are various alternative motives at play in this game. For example, responders might reject for purely distributive reasons (i.e., fairness) and not to punish the proposer, and their decision whether to reject might also be affected by efficiency preferences or altruism. Even if the responder wishes to punish the proposer, though, the proposer's intentions might be obscured by motives other than fairness, such as a self-interested desire to avoid rejection, which is further confounded by risk preferences. Indeed, comparison with other games imply decisions in the ultimatum game result from a confluence of motives, e.g., Forsythe , Horowitz, Savin and Sefton (1994).

discussion around a non-strategic dictator experiment on reciprocal motives. Croson and Konow (2009) introduced a two-stage dictator game in which a dictator first chooses one of six divisions with a recipient of a sum of money, $X = 10$, viz., {(10-0), (8-2), (6-4), (4-6), (2-8), (0-10)}. Then, there follows a previously unannounced second stage, in which a different subject chooses a division between the same subjects of an additional sum of money, $Z = 20$, in any integer amounts. In one experimental condition, the second stage dictator and recipient are the first stage recipient and dictator, respectively. For clarity, I will refer to them consistently according to their first stage roles as D and R, respectively. In this condition, R is in the second stage a *stakeholder*, or party to the allocations. In another condition, the second stage allocator is a third party, or *spectator*, who is paid a fixed sum to allocate *Z* between D and R. Another pair of treatments is identical to these two with stakeholder and spectator versions, except the first stage allocation is not chosen by anyone but rather is randomly assigned, so this is a 2×2 between-subjects design. The strategy method is employed for second stage allocations: all second stage allocators choose a division of *Z* for each of the six possible first stage divisions. The variable of interest is the allocation decisions about *Z* by the second stage allocators based on whether they were themselves a stakeholder or spectator and on whether the first stage division was chosen by a dictator or randomly determined.

Take first the case of the stakeholder, R, who is endowed with the amount received from the first stage, *x*, plus the second stage sum, *Z*, and can transfer any amount, $z \in [0, Z]$, to D (note *f* refers here to D as the second stage recipient). This R's utility function can be written

$$U = u(x + Z - z) + \sigma\phi f(z - \eta_z - \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot x') + \sigma g(z, \alpha)$$

where the entitlement of the current recipient is $\eta_z = {}^{Z}/_{2} - X + x + \eta_x$. That is, D is entitled to one-half of the current endowment, ${}^{Z}/_{2}$, since this is a simple D game. In addition, η_z reverses any inequity in how much D took in the first stage, $X - x - \eta_x$, where $\eta_x = {}^{X}/_{2}$ is the fair division of the first stage stakes, again taken to be equal splits, since this is a simple D game with fixed stakes. We can also specify virtue preferences more precisely for this experiment. As explained in section 2, threshold generosity, $\tilde{\gamma}$, is the break-even level of generosity for reward or punishment and depends on moral salience in the reference state, $\tilde{\sigma}$. But since the choice of reference state is arbitrary and salience depends on many aspects of moral and non-moral context, we can define the reference state to correspond to one in which moral salience is at the

same level as in the first stage of the dictator game at hand. In that case, threshold generosity, $\tilde{\gamma}$, equals the threshold transfer of D in the first stage, $\tilde{x}$. Similarly, R's estimate of D's generosity, $\hat{\gamma}$, then corresponds to D's actual generosity in the first stage, x.[6] Finally, the scale is the fair transfer from D to R, i.e., the fair division of the first stage sum, $x' = {}^{X}/_{2}$, since R is unendowed in this design. Then, R's utility function in the second stage can be written

$$U = u(x + Z - z) + \sigma\phi f\left(z - {}^{Z}/_{2} + {}^{X}/_{2} - x - \theta r(x - \tilde{x}) \cdot {}^{X}/_{2}\right) + \sigma g(z, \alpha).$$

Now we come to the following theorem about stakeholders in this experiment.

THEOREM 3.1: In the two-stage dictator game, second stage allocators, who are stakeholders, partially adjust for first stage transfers that are random, i.e., $0 < {}^{dz}/_{dx} < 1$, but some stakeholders sanction first stage dictators, i.e., ${}^{dz}/_{dx}$ increases, when those dictators choose first stage allocations.

PROOF: See Appendix 1.

When first stage endowments are random, first stage dictators cannot reveal their moral character, so $r = 0$, $\tilde{z} = 0$, and the partial adjustment of z to x follows from allocative preferences. When first stage dictators reveal their character through their choices, however, $r' > 0$, and second stage allocators sanction. These claims are all consistent with the evidence on stakeholder decisions from this experiment.

Note that one part of SF 3.1 that is not claimed in Theorem 3.1 is asymmetric sanctioning. Stakeholder allocations do not produce a clear measure of this asymmetry, given the confluence of additional motives, including material self-interest, inequity aversion and altruism, so we turn now to spectators, whose decisions are predicted to generate a measure that is not distorted by these factors. The utility function of a spectator in this experiment can be written

$$U = u(\bar{z}) + \sigma\phi f\left(z - {}^{Z}/_{2} + {}^{X}/_{2} - x - \theta r(x - \tilde{x}) \cdot {}^{X}/_{2}\right).$$

where $\bar{z}$ represents the fixed payment the spectator receives for making this decision. Remember that no altruism term is included in a spectator's utility function, since the relationship is impartial rather than personal.[7] Theorem 3.2 follows.

---

[6] To be exact, $\hat{\gamma} = x$ for interior solutions, but further specification of $\hat{\gamma}$ is needed in the case of corner solutions. This refinement is unnecessary for the current focus on mean behavior, but it will be addressed in section 4, where it provides insight into an additional finding that becomes apparent in the design discussed there.

[7] Note also that the fairness preference is formulated with respect to the first stage dictator, because that is the

THEOREM 3.2: In the two-stage dictator game with randomly assigned first stage allocations, second stage allocators, who are spectators, equalize, adjusting completely for first stage transfers that are random, i.e., $^{dz}/_{dx} = 1$. When dictators choose first stage allocations, some spectators sanction them, i.e., $^{dz}/_{dx} > 1$, and some equalize. Those who sanction have different thresholds, $\tilde{x}$, and sanction, on average, asymmetrically, punishing more strongly than they reward.

PROOF: See Appendix 1.

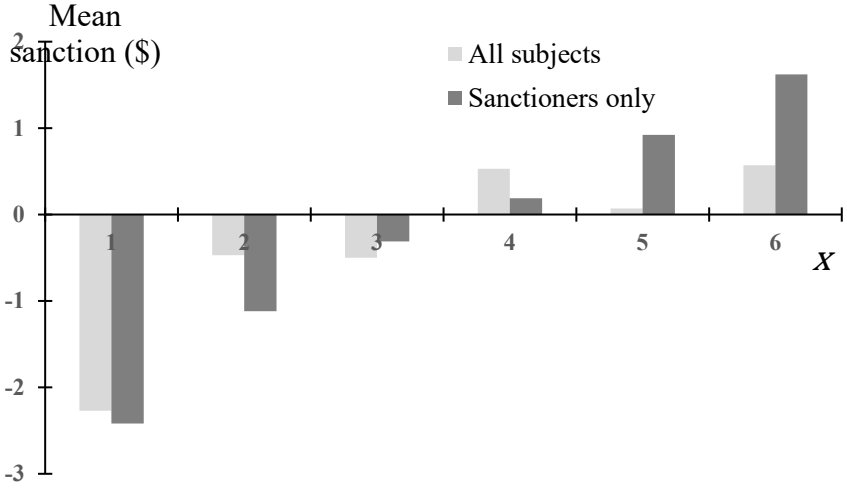

FIGURE 1. – Sanctioning in Croson and Konow (2009).

Spectator transfers are assumed to be motivated solely by fairness and virtue preferences. When first stage allocations are random, spectators adjust *z* to *x* one-for-one, but, when they are chosen by first stage dictators, virtue preferences kick in, and spectators sanction. The asymmetry follows from the fact that $^{d^2z}/_{dx^2} < 0$ due to the concavity of *r*. All of these claims are consistent with the spectator decisions in this experiment. For each level of transfers chosen by first stage dictators, Figure 1 illustrates spectator reward or punishment as the mean amounts by which their transfers to the first stage dictator, *z*, fall short of or exceed equalizing transfers. The mean sanctions of all spectators are illustrated in the light bars, but only 50% of spectators sanction, consistent with the assumption that some agents place zero weight on sanctioning

subject, who has revealed something about his character. A term could be added for fairness toward the first stage recipient, but unfairness toward one first stage subject is simply mirrored by unfairness in the opposite direction toward another, and the conclusions are qualitatively unaffected, so I avoid this clutter as in previous analysis of spectator allocations in dictator games, e.g., Konow (2000).

($\Theta(0) > 0$). Another 37% equalize, and the other 13% cannot be put into either category.[8] The mean reward and punishment of only those who sanction are illustrated in the darker bars. Figure 1 suggests an asymmetry, which is corroborated in more formal analysis in Croson and Konow and is generally consistent with third party sanctioning in other studies, such as Almenberg et al. (2011) and Fehr and Fischbacher (2004). Moreover, the results reveal the assumed heterogeneity in thresholds: of spectators who sanction, 27% stop punishing (and thereafter either equalize or begin rewarding) at a first stage transfer of 0, 40% at 2, 20% at 4, and 13% at 6.

4. Sinners and Saints

This section introduces an experiment that provides an out-of-sample test of the theory presented in this paper while shedding light on other findings, including the taking effect. In what I will call the "sinners and saints" game, there is a first stage in which dictators may give to or take an amount x from the endowments of the dictators and recipients, where $X > Y > 0$. The endowments are always fixed at the same level, but the range of permissible transfers varies across "*cases*." Cases are varied between subjects, whereby, for a given case, the minimum possible transfer, i.e., the most the D can take, is denoted $x^L \leq 0$, and the maximum possible transfer is denoted $x^H > 0$. Then, there is an unannounced second stage in which a spectator is paid a fixed amount, $\bar{z}$, to allocate an additional larger sum, $Z > X + Y$, between the D and the R in the first stage, which is contingent on each possible D choice for x, i.e., the strategy method is used. To employ a metaphysical conceit, an agent decides how to treat a patient during his mortal life, ignorant of the afterlife in which an impartial judge metes out sanctions on sinners and saints that involve even higher stakes. This experiment, while novel, merges design features that have been well validated elsewhere and that, therefore, relate to a broader set of results.[9]

The first stage of the sinners and saints game is a variation on a dictator game with taking, so I begin with a brief review of the stakeholder analysis, including assumptions, stylized

[8] The comments of the 13% in the post-experimental questionnaire indicate that half misunderstood their task and that the other half believed (mistakenly) that there was some strategic dimension of the decisions.
[9] Spectators in a two-stage dictator experiment were introduced in Croson and Konow (2009) and discussed in the prior section, but the design of that study differed in other ways from the present one: there are never any taking options, in treatments where Ds can choose, only Ds are endowed but not Rs, in other treatments, endowments are random and Ds cannot transfer, and the range of permissible transfers is not varied. As with the current design, Krupka and Weber (2013) used spectators and stakeholders to analyze taking, but their design differs in that third parties are not used to sanction but rather to provide self-reported appropriateness ratings of stakeholder transfers on a point scale, stakeholder endowments are varied, and taking options are not varied.

facts, and theoretical conclusions, which are numbered here as in Konow (2022), where they are presented in greater detail. A motivation for the taking effect can be seen in otherwise law-abiding citizens, who sometimes join in looting during civil disturbances and natural disasters. Dictator experiments with take options demonstrate that such taking behavior persists, even after controlling for possible extrinsic incentives, such as the fear of punishment. The taking effect is an example of a class of anomalies involving variation in helping and harming opportunities.[10] In a dictator game where Ds and Rs are endowed $X > Y > 0$, assume giving opportunities constitute moral context and taking opportunities non-moral context. Specifically, for concreteness, let the moral measure be $m(C_i) = \max\{c_i \in C_i\} - \min\{c_i \in C_i\}$, $C_i = \{C_+, C_-\}$, where C_+ is the set of non-negative transfers from D to R and C_- the set of negative transfers, i.e., transfers from R to D (Assumption 7). Then, adding take options reduces giving on both the intensive and extensive margins, i.e., the mean transfer and frequency of positive transfers decrease, and mean transfers fall at a decreasing rate (SF 6.1 and Theorem 6.1).

The focus of most of the theoretical and empirical analysis here is on second stage spectators, who allocate an additional larger sum, Z, between the first stage D and R as with the two-stage dictator game of the prior section. Specifically, the utility function of the spectator in the sinners and saints game is

$$U = u(\bar{z}) + \sigma\phi f\left(z - \eta_z - \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot {}^{X}/_{2}\right),$$

where moral preferences are formulated with respect to the D from the first stage, whose choice provides a signal of the D's character. Thus, z denotes the amount of Z that the spectator allocates to D with the remainder of $Z - z$ going to R, η_z is the D's entitlement in the second stage, $\hat{\gamma}$ is D's estimated notional generosity to R, and $\tilde{\gamma}$ is the spectator's threshold for rewarding or punishing D. The usual assumption that fairness reduces to equal splits in a simple D game like this implies that $\eta_z = {}^{Z}/_{2} - {}^{X}/_{2} - {}^{Y}/_{2} + x$, i.e., η_z calls for equal splits of the total endowments and corrects for any shortfall or excess vis-à-vis equality in D's first stage transfer.

The following theorem states several predictions for this experiment.

Theorem 4.1: In the sinners and saints game, second stage spectators sanction, and sanctions are concave in dictator first stage transfers. There is a discontinuous increase in reward (or

[10] Another example of this class of anomalies, which is discussed in Konow (2022), is joy-of-destruction, where agents may, with no personal gain, destroy others' earnings.

decrease in punishment) at x^H, and a converse discontinuity at x^L, i.e., a decrease in reward (or increase in punishment). Holding x^H constant, increasing dictator taking options, i.e., lowering x^L, implies a lower threshold for spectator sanctioning, $\tilde{x}$, and also increases reward, or decreases punishment, of dictators by spectators at every level of dictator transfers.

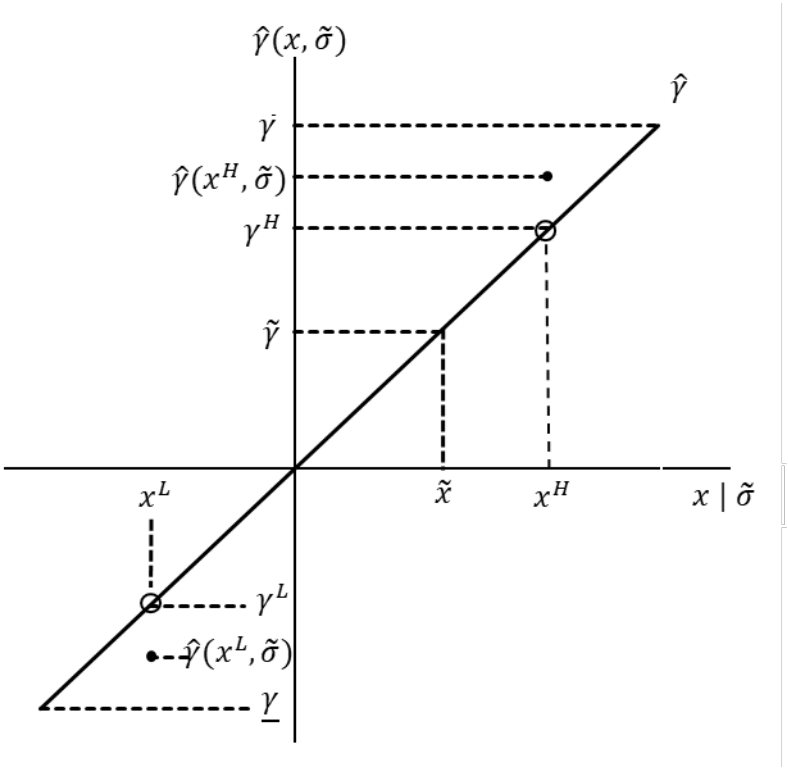

FIGURE 2. – Notional and effective generosity.

The proof of this theorem can be found in Appendix 1, but the reasoning is illustrated in Figures 2 and 3. First, the context of the reference state may be defined to match the salience in the first stage of this dictator game, $\tilde{\sigma}$. In a game with the same set of permissible transfers as in the reference state, expected generosity, $\hat{\gamma}$, equals notional generosity, γ, both of which equal the dictator's transfer, x, for the full range of dictator types from the least generous, $\underline{\gamma}$, which is the greatest lower bound of notional generosity, to the most generous, $\bar{\gamma}$, which is the lowest upper bound of notional generosity. This is illustrated in Figure 1 by the 45-degree line. Thus, we can write notional generosity, $\gamma(x, \tilde{\sigma})$, as a function of x and $\tilde{\sigma}$. As noted in section 2, however, notional generosity may differ from effective generosity, x, because of constraints on choices. Suppose at least some dictators are constrained to giving less than their preferred amount, i.e., $x^H < \bar{\gamma}$, and/or from taking less than their preferred amount, i.e., $x^L > \underline{\gamma}$. Due to this censoring,

a spectator's estimate of the notional generosity of a dictator who chooses x^H, $\hat{\gamma}(x^H, \tilde{\sigma})$, is greater than that of the dictator type, who would choose x^H in the reference state, $\gamma^H = \gamma(x^H, \tilde{\sigma})$, since it includes not only those who notionally prefer x^H but also others who prefer a larger transfer but are prevented from transferring it. Similarly, the spectator's estimate of the generosity of a dictator who chooses x^L, $\hat{\gamma}(x^L, \tilde{\sigma})$, is less than that of the dictator type, who notionally prefers x^L, $\gamma^L = \gamma(x^L, \tilde{\sigma})$. Suppose that the spectator's character threshold for dictator generosity is $\tilde{\gamma}$ and that $\gamma^L < \tilde{\gamma} < \gamma^H$. Then, in the interior, $\tilde{\gamma} = \tilde{x}$, and the spectator rewards all $x > \tilde{x}$ and punishes all $x < \tilde{x}$, and, generalizing Theorem 3.2, sanctions are asymmetric due to the concavity of r in x. One exception to concavity is occasioned by the censoring of γ at x^H: the optimal sanction, $\tilde{z}$, is increasing in γ, so the discontinuity here implies a discontinuous increase in reward (or reduction in punishment). Because of the predicted discontinuities at x^L and x^H, these will be treated separately in the regression analysis later.

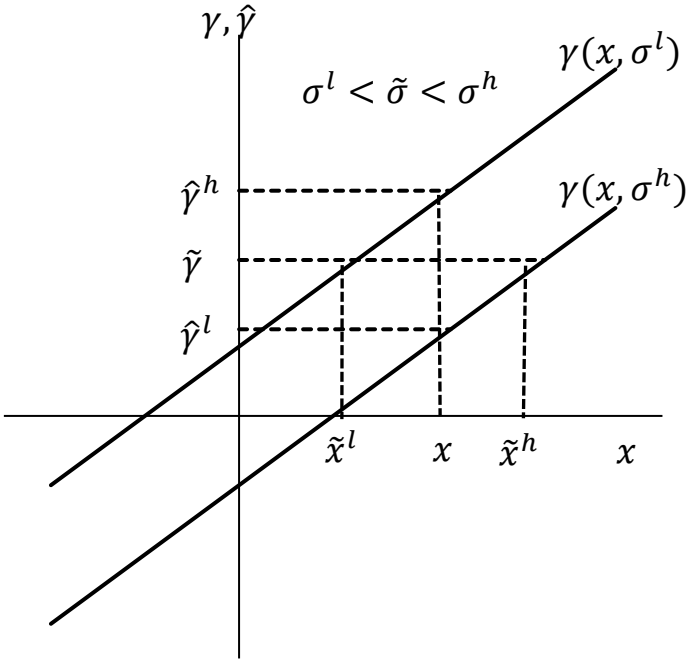

FIGURE 3. – Generosity and moral salience.

Other claims of Theorem 4.1 are illustrated in Figure 3, which focuses on interior solutions. Estimated (and notional) generosity can be written as a function of x and σ, i.e., $\hat{\gamma}(x, \sigma)$. The main analysis involves variation in the amounts that may be taken, which, as already discussed, affects moral salience. Starting from the reference level of salience, $\tilde{\sigma}$, consider an increase in taking options (the line going through the origin that corresponds to the reference state is omitted here to avoid clutter). Ceteris paribus, salience falls to σ^l, $\sigma^l < \tilde{\sigma}$, and

a dictator, who would make a transfer, say, equal to the spectator's threshold of $\tilde{\gamma}$ under $\tilde{\sigma}$, will now give less, $\tilde{x}^l$. That is, the schedule representing the dictator's notional generosity shifts to the left, $\gamma(x, \sigma^l)$, and the spectator's threshold for sanctioning falls. Similarly, a reduction in taking options increases salience to σ^h, $\sigma^h > \tilde{\sigma}$, and the same dictator will now give more, $\tilde{x}^h$, shifting the schedule to the right, $\gamma(x, \sigma^h)$, and increasing the spectator's threshold for sanctioning. Finally, changes in taking options also affect the spectator's estimate of a dictator's type and, therefore, the spectator's sanctioning of the dictator (which is the same as the dictator's notional type for interior solutions). A dictator, who gives x under low salience, σ^l, reveals higher intrinsic generosity, $\hat{\gamma}^h$, than one, who gives the same amount under high salience, σ^h, and reveals lower generosity, $\hat{\gamma}^l$. Thus, the same transfer may be praiseworthy in the former case but blameworthy in the latter, according to spectators. The final claim of Theorem 4.1 concerns the size of the discontinuous increase in z at x^H and the relationship to x^L. An expansion of x^L reduces σ, censors fewer dictator types, and increases $\hat{\gamma}(x^H, \sigma)$. This reduces the discontinuity, if the increase in $\hat{\gamma}$ is smaller than the change in censored types.

The parameters of this experiment are as follows: $X = 15$, $Y = 5$, $Z = 40$, and $x^H = 5$ for all of the three main cases, which differ according to the values of x^L, viz., the Give 5 case with x^L=0, Take 1 case where $x^L = -1$, and Take 5 case where $x^L = -5$. Note that $x^H = 5$ allows X to equalize payments between the X and Y. The values for x^L were chosen to permit testing of the theoretical predictions and to allow comparisons with prior taking games. The main treatment of the sinners and saints game that has been described thus far is called the Double dictator treatment: subject X chooses how much to transfer to or from subject Y out of their aggregate 20 points, and subject Z is paid a fixed \$5 to allocate 40 points between subjects X and Y for each possible transfer by X to Y, where one point always equals \$0.20. In order to examine whether the entitlement changes with cases, there is also a so-called Benevolent dictator treatment in which X and Y are endowed as in the other treatment, i.e., $X = 15$ and $Y = 5$, and each Z subject is paid a fixed \$2 to choose a transfer of these first stage points between X and Y subjects. Thus, the design is the same as the first stage of the Double dictator treatment with the same cases, except Z instead of X chooses transfers between X and Y, and there is no second stage decision. To avoid any spillover effects of roles and decision contexts on allocations, all decisions were collected between subjects, i.e., different subjects were used in the roles of X, Y and Z, in the two treatments, and in the different cases. The complete protocol can be found in

Appendix 2 (not for publication).

	Treatment	
Case	Double dictator	Benevolent dictator
Give 5	**66 X**, 66 Y, **66 Z**	30 X, 30 Y, **30 Z**
Take 1	**62 X**, 62 Y, **62 Z**	45 X, 45 Y, **45 Z**
Take 5	**63 X**, 63 Y, **63 Z**	37 X, 37 Y, **37 Z**

TABLE 1
SINNERS AND SAINTS DESIGN

The experiment was programmed in oTree (Chen, Schonger and Wickens, 2016) and conducted on Amazon MTurk. Table 1 illustrates the experimental design, showing for each case and treatment the number of participants in each role, whereby those in decision-making roles are denoted in bold font. Similar to many related studies (e.g., Bardsley, 2008, Chowdury, Jeon, and Saha, 2017, Grossman and Eckel, 2015, Korenok, Millner and Razzolini, 2012), a minimum of roughly 30 observations (i.e., 30 triples) per case were targeted for the Benevolent dictator treatment, and a minimum of twice that number, viz., 60 triples per case, were targeted for the Double dictator treatment, since it is the main treatment of interest. The actual numbers usually exceed these minimums due to differences in the timing of when subjects were cut off from entering. For this study, an MTurk subject pool was preferred for a number of reasons. A substantial literature now exists that MTurk participants behave similarly to university student subjects in qualitative terms. Moreover, MTurkers are typically closer to the general population in terms of demographic characteristics and average experimental behavior, e.g., Snowberg and Yariv (2021) find the average generosity of MTurk dictators intermediate to that of the more selfish students and that of a more generous representative sample.[11] In addition, the total sample size desired for this study was larger than that accessible at any given time from most student subject pools (the results are based on a total of 1029 participants). Moreover, this study lends itself to the adoption of measures to address typical concerns about an online subject pool (e.g.,

[11] Johnson and Ryan (2019) conclude that quality is not harmed by the lack of control and lower stakes on MTurk. Moreover, the equivalency of results from student and MTurk subjects extends to designs involving moral preferences, such as prisoner's dilemmas (e.g., Horton, Rand and Zeckhauser, 2011), public goods games (e.g., Arechar, Gächter, and Molleman, 2018), and dictator games (e.g., Snowberg and Yariv, 2021).

see Hauser, Paolacci, and Chandler, 2019).[12] Including the $2 show-up fee (called a reward in MTurk), the average earnings were $6.25 for an average of 20-25 minutes of most subjects' time. This is several times the usual MTurk pay of $1-$5 per hour and similar to the hourly wages used by Snowberg and Yariv, moreover, they report their results are robust halving the incentives.

	Mean transfer (SD)	Positive transfers (%)
Give 5	3.45 (2.105)	78.8
Take 1	2.16 (2.343)	64.5
Take 5	1.78 (3.777)	61.9

TABLE 2
TRANSFERS BY X TO Y IN DOUBLE DICTATOR TREATMENT

Turning now to the results, Table 2 summarizes the transfers of X subjects in the Double dictator treatment. As predicted, the mean transfers and percentage of positive transfers decrease, as x^L falls. According to two-sided t-tests tests of differences in means, the mean transfer is lower in the Take 1 (p=0.001) and in the Take 5 (p=0.002) cases than in the Give 5 case, but Take 1 does not differ significantly from Take 5 (p=0.497). Compared to the Give 5 case, two-sided z-tests tests of proportions show a decrease in positive transfers that is marginally significant in the Take 1 case (p=0.073) and significant at conventional levels in the Take 5 case (p=0.036), but Take 1 and Take 5 do not differ significantly (p=0.762). These findings are similar in direction and significance to prior related taking games except that these X subjects are, on average, more generous and less likely to take, when given the opportunity. This is consistent with the expectation that the more representative sample here is more generous than student subjects, who were used in prior studies. A second contributing factor is surely the relatively high ratio of X to Y endowments of 3:1 here versus the lower ones (usually 2:1) used

[12] To address concerns about English language fluency, participation was restricted to US residents. Numeracy was established with a test consisting of three fill-in-the-blank questions on addition and subtraction that permitted at most two attempts each before disqualification. To address concerns about attention and comprehension, subjects had to complete correctly within two attempts each of three questions in a quiz about the instructions (two quizzes of three questions each in the case of Z subjects in the more complicated Double dictator treatment). To minimize attrition, each subject faced only one type of decision and the non-strategic design permitted non-simultaneous collection so that subjects did not have to wait for other subjects. The simple design helped keep the study short and address both subject attention and attrition: subjects were permitted up to one hour, but most completed it in less than thirty minutes. Self-selection biases are presumably less problematic with MTurk than with a university subject pool, but that concern was further addressed by describing the study in general terms as an "Academic experiment involving decisions about the distribution of money." As an aside, the data collection took place during the 2020-21 COVID pandemic at a time when laboratory experiments were not feasible, but that fact had no bearing on the choice of an online format, which had been previously planned based on its advantages for this experiment.

in prior studies, including Bardsley (2008), List (2007), and Zhang and Ortmann (2013).

Having established that X decisions are consistent with theoretical predictions for stakeholders and qualitatively replicate prior findings, we turn now to the spectator Z allocations in the Double dictator treatment. Table 3 provides a summary of Z transfers to X subjects. The first row presents the full range of X transfers considered and the second row the corresponding Z transfers to X that equalize total earnings between X and Y. The remaining rows report the mean allocation by Z to X, z, for each level of X transfer, x, for each case, which permits a preliminary impression of the results. As predicted, z is monotonically increasing in x within each case in every instance and monotonically decreasing in x^L for each given value of x save in one instance (Take 1, $x = 5$). Comparison of these means with the equalizing Z allocations suggest that Zs usually punish Xs, on average (shaded in red), for transferring less than the 5 points that equalize first stage payoffs: z exceeds the value that equalizes total earnings (shaded in blue), only for an x of 4 or 5 in the Take 5 case and for an x of 5 in the Take 1 case. This is consistent with the predicted shift in $\tilde{x}$ with x^L: the change from Take 5 to Take 1 to Give 5 represents a progressive increase in x^L and in salience and, therefore, an increase in the threshold for rewarding X transfers.

X transfer	-5	-4	-3	-2	-1	0	1	2	3	4	5
Equalizing Z allocation	10	11	12	13	14	15	16	17	18	19	20
Take 5	5.17	5.89	6.03	7.13	8.02	13.08	15.27	16.19	17.92	19.40	20.70
Take 1					6.10	9.53	12.87	15.85	17.73	18.65	21.71
Give 5						7.82	11.44	12.82	13.68	16.03	19.08

TABLE 3
MEAN ALLOCATIONS BY Z SUBJECTS TO X SUBJECTS IN DOUBLE DICTATOR TREATMENT

Key: Red (blue) allocations are below (above) equalizing ones. Means differ from equalizing transfers according to t-tests at the 5%/10% level of significance; lightly shaded results are not significant at conventional levels.

Figure 4 presents a scatterplot of Z decisions, where circle sizes are proportionate to the frequency of each choice and colors correspond to cases: yellow for Give 5, green for Take 1 and blue for Take 5. The main patterns are consistent with theory: z is increasing in x, on average, and some Z subjects equalize, indicated by larger circles along a diagonal, whereas others sanction, reflected mostly by punishment below the diagonal. Turning to multivariate analysis of Z decisions, consider the following regression equation:

$$z_{it} = \alpha + \beta_x \ln(x_t + 2) + \beta_1 Take1 + \beta_5 Take5 + \beta_L x^L + \beta_H x^H + \varepsilon_i$$

where z_{it} is the allocation of Z subject i to subject X based an X transfer of x_t, α is the constant,

the βs are the coefficients on the independent variables, and ε_i is the error term. The omitted case is Give 5, so Take1 and Take5 are dummy variables for those cases, respectively. Discontinuities are predicted at x^L and x^H, so dummy variables are included at the lowest possible transfer in each respective case for x^L and at $x^H = 5$, where x^H is common to all cases. The analysis focuses on Tobit regressions with left-censoring using a logarithmic specification of x_t for values between -1 and 5, which is why 2 is added to x_t, making the minimum value of this independent variable 1. There are several reasons for these choices. First, the non-linear specification is consistent both with theoretical expectations and with the results of regression estimations illustrated by the logarithmic trendlines in Figure 4. Second, the fact that second stage allocations are twice as large as the stakes in the first stage (viz., 40 vs. 20 points) is a feature designed to give wide berth to second stage allocators, and inspection of the scatterplot suggests this was largely successful. Nevertheless, scatterplots also reveal considerable censoring of z at values of x, especially below -1, and recommend the use of Tobit. Third, the chief interest is in comparison of differences across cases where there are common values of x for at least some cases. Moreover, including additional negative x values for the Take 5 case produces results that are qualitatively similar but risks producing estimates that give disproportionate weight to the Take 5 case and its increasingly censored values (reaching 46% of allocations when x equals -5).

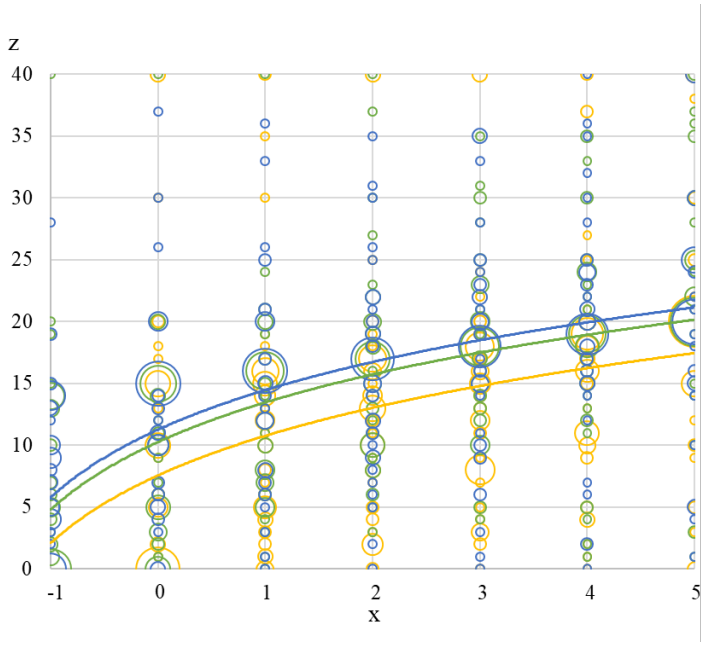

FIGURE 4. – Scatterplot and logarithmic regression trendlines of Z allocations to X (z).
Key: Yellow: Give 5, Green: Take 1, Blue: Take 5

Table 4 presents the results of regression analysis, whereby standard errors are clustered at the level of the 191 individual Z subjects. Column 1 shows the results of an OLS regression and column 2 the results of a Tobit regression with left-censoring. The results are qualitatively the same except for the x^L dummy variable, which is negative in both estimations but turns significant when account is taken of the left-censoring that is compromising the OLS estimation. We focus, therefore, on regression (2), the results of which are all consistent with Theorem 4.1, some at high levels of significance. Spectator sanctions are increasing and concave in X transfers to Y. As taking options expand, spectators progressively increase reward, or decrease punishment, significantly in the Take 1 and Take 5 treatments. The coefficient for the x^L dummy confirms the predicted discontinuous decrease in z and the coefficient for the x^H dummy the predicted discontinuous increase in z.

	(1) OLS	(2) Tobit
$\ln(x+2)$	6.516***	6.822***
	(0.509)	(0.352)
Take 1	2.507*	2.662*
	(1.105)	(1.112)
Take 5	3.587**	3.680***
	(1.169)	(1.114)
x^L	−0.580	−1.533*
	(0.749)	(0.636)
x^H	1.769***	1.673***
	(0.417)	(0.464)
Constant	4.021***	3.398***
	(1.188)	(0.941)
R^2	0.255	

TABLE 4

REGRESSION ANALYSIS OF Z ALLOCATIONS TO X

N=191. Tobit regressions are left-censored. Standards errors are clustered at the individual subject level and are reported in parentheses. */**/*** denotes significance at the 5/1/0.1-percent level.

The theory advanced here has been shown to be highly consistent with the results on stakeholder and spectator decisions in the sinners and saints game. As always in such instances, that fact does not rule out the possibility of alternative explanations, such as those offered by the reciprocity theories mentioned in section 3. For example, Rabin (1993) and Dufwenberg and Kirchsteiger (2004) define fairness as the average of the highest and lowest efficient payoffs, which implies fair allocations (η) vary directly with x^L. Falk and Fischbacher (2006) define

fairness as equal payoffs, but their theory generates equivalent predictions about the effect of variation in x^L by building the effect of the choice set into their "intention factor." Reciprocity theories are typically formulated for stakeholders, but, for simplicity, I will cast them in the current spectator framework and analyze the effect they predict of x^L through η.

Theorem 4.2: In the sinners and saints game, let the fair allocation in the first stage and the threshold for sanctioning be functions of the minimum permissible transfer, i.e., $\eta_x(x^L)$ and $\tilde{x}(x^L)$, respectively, where $\frac{\partial \eta_x}{\partial x^L} > 0$ and $\frac{\partial \tilde{x}}{\partial x^L} > 0$. Then lowering x^L increases reward, or decreases punishment, at every level of dictator transfers.

Proof: See Appendix 1.

Reciprocity theories provide a partial account for observed spectator sanctioning: they are consistent with the observed shift in the threshold and the level of sanctioning. They do not, however, predict the discontinuities we observe at x^L and x^H. Moreover, reciprocity theories and the theory proposed here diverge in their predictions for behavior in the aforementioned Benevolent dictator treatment, in which a spectator chooses transfers between X and Y. In this treatment, the spectator's utility function is simply

$$U = u(\bar{z}) + \sigma\phi f(x - \eta_x),$$

which yields the following predictions.

Theorem 4.3: In the Benevolent dictator treatment, the spectator allocates to subjects their entitlements. That means the spectator's allocation does not vary with x^L according to the theory of moral salience, conditional altruism and virtue preferences, but it does vary directly with x^L, according to reciprocity theories.

Proof: By the first order condition, $\sigma\phi \cdot {}^{\partial f}/_{\partial w} = 0$, $w = x - \eta_x$, which implies $x = \eta_x$. In this game, η_x is fixed in the theory advanced here, whereas $\frac{\partial \eta_x}{\partial x^L} > 0$ in reciprocity theories.

Table 5 presents the results of this treatment for the three main cases as well as for an additional case, Give 10, which I will discuss momentarily. For the three main cases, the mean transfers range from \$4.04 to \$4.47. The three pairwise tests reported in the table reveal that none of these differences is significant even at the 20% level, whereby all p-values in this table are two-sided. This is consistent with the theory proposed here but not with reciprocity theories. A different question concerns spectator estimates of the fair transfer. Recall, as stated in section 2, that the entitlement is assumed to reduce to equal splits in simple games with fixed stakes but

that theoretical claims do not generally depend on this specific value. Equality implies an X transfer of 5, in this case, but the tests reported in Table 5 show that the means in the three main cases all differ significantly from 5. Nevertheless, while comparisons of mean spectator transfers provide valid conclusions about whether the entitlement varies with x^L, they might not provide good estimates of the entitlement itself. As with all experiments, subject decisions here are noisy, but variance is censored on the right at the value of 5 in the three main cases, which creates a downward bias in the estimate of η_x. For that reason, the benevolent dictator treatment includes an additional case, Give 10, in which spectators may allocate any amount from 0 to 10 to subjects X. For this case, the mean transfer is $4.68, which does not differ significantly from 5.

			Difference in Means p-values (t-statistics)			
	Mean transfer (SD)	*N*	$H_0: \eta_x = 5$	Give 5	Take 1	Take 5
Give 5	4.47 (1.335)	30	0.034 (2.175)			
Take 1	4.04 (1.673)	45	0.000 (3.849)	0.242 (-1.179)		
Take 5	4.41 (1.077)	37	0.001 (-3.332)	0.839 (-0.204)	0.249 (1.161)	
Give 10	4.68 (1.738)	40	0.245 (1.165)	0.584 (0.551)	0.088 (1.729)	0.420 (0.812)

TABLE 5
TRANSFERS BY Z TO Y IN BENEVOLENT DICTATOR TREATMENT

This section introduced an experiment designed as an out-of-sample test of the theory proposed here. The results of the experiment are uniformly consistent with the predictions of the theory and, in several respects, inconsistent with reciprocity theories. In the next three sections, the theory is applied to additional decision contexts.

5. Outcome Bias

Numerous studies across multiple disciplines have established a preference for rewarding or punishing individuals based on uncontrollable (or brute) luck, including in politics (e.g., Healy, Malhotra, and Mo, 2010), sports (e.g., Kausel, Ventura and Rodriguez, 2019), and CEO compensation (e.g., Bertrand and Mullainathan, 2001). In the law, someone, who kills another accidentally, can be sentenced more harshly than someone, who meant to kill another but failed (Cushman et al., 2009). In economics, this so-called *outcome bias* has also been extensively

studied experimentally, especially, in the context of the principal-agent problem (e.g., Charness and Levine, 2007, Rubin and Sheremeta, 2016). Although the optimal contract rewards effort and disregards luck, people sanction luck, even when decision makers are clearly not responsible (Gurdal, Miller and Rustichini, 2013) and even when those who are sanctioning are third parties (Brownback and Kuhn, 2019). Moreover, outcome bias is enigmatic from the perspectives of philosophy (e.g., Williams, 1981) and of reciprocity theories in behavioral economics, which conceptualize reward and punishment in terms of intended consequences (e.g., Rabin, 1993, Falk and Fischbacher, 2006).

Although outcome bias is inconsistent with optimal contracts and reciprocity theories, I argue that it is consistent with moral intuition and virtue preferences. Indeed, in a more general specification of virtue preferences, outcome bias is not a bias, at all. Consider the intuition: two individuals, who are operating a motor vehicle, run a stop sign, the one without consequence and the other causing the death of a family. Do we legally, and should we morally, really hold the two equally accountable? In the one case, the driver might pay a fine of a few hundred dollars, whereas, in the other case, the driver can be found guilty of manslaughter and serve jail time. I claim that "outcome bias" is a feature, not a bug, of moral preferences: in virtue preferences, the relevant sanctioning motive is with respect to intent coupled with consequential action. Specifically, the difference in sanctions reflects the intuition exemplified above, which can be incorporated by scaling virtue preferences according to whether or not the intended outcome obtained. Consider the following stylized facts from economics experiments on this topic.

SF 5.1: When an agent can choose actions with uncertain outcomes, others sanction the agent, meaning these rewards and punishments are not explained by distributive preferences alone. Sanctions are asymmetric: low generosity is punished more strongly than high generosity of equal magnitude (e.g., Cushman et al., 2009, de Oliveira, Smith and Spraggon, 2017).

SF 5.2: Under uncertainty, sanctions are based not only on the chosen action (and its expected outcome) but also the realized outcome, for which the agent is not responsible. The sanctions for actions leading to outcomes that are both expected and realized are greater than those for outcomes that are expected but not realized (Charness and Levine, 2007, Gurdal, Miller and Rustichini, 2013, Rubin and Sheremeta, 2016), and both stakeholders as well as third parties exhibit this behavior (Brownback and Kuhn, 2019, Gino, Shu and

Bazerman, 2010, Sezer, Zhang, Gino and Bazerman, 2016). The sanctions for realized outcomes that agree with expected outcomes are roughly the same as those for the same outcomes chosen with certainty (Cushman et al., 2009).

Despite considerable variation in features of these experiments, including in the role of uncertainty, effort, information and payoff functions, the findings are quite consistent. Given these design differences and the focus of the current analysis, therefore, I will analyze a hybrid design that captures elements of different studies and fits the focus here on non-strategic decisions. In what I will call the "fair luck game," there are two stages, whereby first stage Ds select an option from a discrete set of risky choices that differ in their expected fairness, and then, in an unannounced second stage, the Rs may sanction the Ds based on the latter's choices and the realized payoffs. Specifically, suppose the first stage payoffs to D,R can be either "fair" (F,F) or "unfair" (H,L), where $0 \leq L < F < H$ and $L + H = 2F = X$. The first stage D makes a risky choice, $x \in \{f, u\}$, involving probability $q > 0.5$, which results in expected payoffs to R, EX^f and EX^u, respectively, of

$$EX^f = qF + (1-q)L$$

$$EX^u = (1-q)F + qL$$

where, obviously, $EX^f > EX^u$.[13]

The subject matter of outcome bias is sanctioning, so we focus on the second stage in which the first stage R may sanction the first stage D by adding money to, or deducting money from, D's payoff (I will refer to them consistently as R and D according to their roles in the first stage). This game requires specification of moral character, or generosity in this context, γ, to accommodate decisions under uncertainty. Analogous to deterministic decisions, consider a reference state in which the first stage D may choose a benefit to the R but now let γ denote the expected payoff to R, which is distributed on the interval $[\underline{\gamma}, \overline{\gamma}]$. In addition, suppose R expects

[13] This game is similar to Cushman et al. (2009) except for two design features. First, the risky options in the hybrid number two, as in Sezer et al. (2016), rather than three, both in order to simplify the analysis and because there is empirically little difference between the second and third choices in Cushman et al. Second, in Cushman et al. the possibility of sanctions is common knowledge, but if first stage dictators anticipate sanctions in the second stage, their choices might be distorted by strategic self-interest. Cushman et al. address this concern by making the probability that sanctions are implemented negligible (viz., 0.1). In fact, the differences in punishment are sufficiently small that expected payoffs are not appreciably affected such that the ranking of more or less generous choices should be preserved. Alternately, the potential problem could be obviated by employing a previously unannounced second stage (see Croson and Konow, 2009), which is assumed in the hybrid, since the results are qualitatively the same, but this approach simplifies the formal analysis.

there is a D type, γ^i, who is indifferent between f and u, where $\underline{\gamma} < EX^u < \gamma^i < EX^f < \overline{\gamma}$. Whereas γ represents notional generosity, effective generosity is constrained in this game to a binary choice between EX^u and EX^f. Thus, R's estimate of D's notional generosity, $\hat{\gamma}$, is either $\hat{\gamma}^f$ or $\hat{\gamma}^u$, depending on D's choice of either f or u, which, respectively, equal

$$\hat{\gamma}^f = \int_{\gamma^i}^{\overline{\gamma}} \gamma\, \rho(\gamma) d\gamma \Big/ \int_{\gamma^i}^{\overline{\gamma}} \rho(\gamma) d\gamma$$

and

$$\hat{\gamma}^u = \int_{\underline{\gamma}}^{\gamma^i} \gamma\, \rho(\gamma) d\gamma \Big/ \int_{\underline{\gamma}}^{\gamma^i} \rho(\gamma) d\gamma.$$

Note that $\hat{\gamma}^u < \hat{\gamma}^f$, and it is further assumed that $\hat{\gamma}^u \leq \tilde{\gamma} \leq \hat{\gamma}^f$, i.e., R's threshold for sanctioning D lies within the interval of D's estimated generosity.

Now consider R's payoff in the second stage. In the fair luck game, R receives a fixed amount from the first stage, x_r, i.e., the realization of D's choice in the first stage. Next is the question of the price R pays to sanction D. In the two-stage D game discussed in section 3, R allocates a fixed sum between D and R. That is, letting z be the amount added to or subtracted from D's payoff in the second stage and Y the amount, as a result of z, that is added to or subtracted from R's payoff, then in the two-stage D game, $dY/dz = -1$. The fair luck game is different in that sanctioning is free and produces neither gains nor losses for R, that is, $dY/dz = 0$. Unlike the prior case, then, there are efficiency implications of R's decision. As previously noted, a large volume of research finds evidence that social preferences include, and sometimes are even dominated by, efficiency concerns (e.g., Charness and Rabin, 2002, Engelmann and Strobel, 2004). The results of experiments on sanctioning discussed here are also consistent with the idea that, when agents can sanction and $dY/dz = 0$, efficiency preferences crowd out other allocative preferences (e.g., Bartling et al., 2014, Bartling and Fischbacher, 2012, Cushman et al., 2009).[14] I model this effect with the parameter $\beta = -dY/dz$, whereby, in the cases considered here, $\beta \in [0,1]$. The D's entitlement in the second stage is assumed to be

$$\eta_z = \left(\frac{1+(1-\beta)b}{2}\right) Z - \beta(X - x - \eta_x).$$

Thus, in the prior two-stage D game where $\beta = 1$ and efficiency plays no role, $\eta_z = Z/2 - X + x + \eta_x$, which, as before, splits the second stage sum equally and corrects any inequity from the

[14] In fact, they are so strong in Cushman et al. that roughly one in six second stage allocators transfer the maximum regardless of first stage actions or outcomes.

first stage. In the current fair luck game where $\beta = 0$ and only efficiency matters, $\eta_z = \left(\frac{1+b}{2}\right) Z$, where $0 < b < 1$. The highest possible payoff is Z, which equals zero in games with only punishment, and $\eta_z = 0$. In games with reward, $Z > 0$ and $\frac{1}{2} Z < \eta_z < Z$ because $0 < b < 1$.[15]

The R's utility function in the fair luck game can, therefore, be written

(4) $U = u(x_r) + \sigma\phi f(z - \eta_z - \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot x') + \beta\sigma g(z, \alpha)$.

This reflects R's fixed payment from the first stage. In addition, as stated, when agents sanction and $\beta = 0$, efficiency is assumed to crowd out other allocative preferences, so the effect of β on altruism is technically included but superfluous, in this case. The final step in specifying virtue preferences to accommodate outcome bias involves the scale, x'. So far, choices, if implemented, have had certain consequences in the cases considered, and the scale was defined as the fair transfer to the patient, i.e., the entitlement in the case of an unendowed patient. Now the consequences of choices are uncertain, and outcome bias is a reflection of the dependence of sanctions not only on choices but also outcomes. In the fair luck game, the fair allocation from the first stage sum of X is $\eta_x = F$, so it makes sense for this to be the scale, when intended and realized outcomes align, whether choices are under certainty or uncertainty. In this case, I will write the choice and realized outcome as the pair $(x, x_r) \in \{(f, F), (u, U)\}$. What about the cases, when intended and realized outcomes do not agree, i.e., $(x, x_r) \in \{(f, U), (u, F)\}$? It is natural, in this case, to think in terms of expected outcomes. As attempted murder is not punished as harshly as murder, so also the scale responds to the difference between expected and realized outcomes. And as attempted murder is punished more severely than attempted burglary, so also the scale responds to differences in expected outcomes. I propose defining the scale in these cases as the expected value from the choice, i.e., EX^f if the choice is fair, and EX^u if the choice is unfair. Then, the scale of sanctions can be defined as

$$x' = \begin{cases} F \text{ if } (x, x_r) \in \{(f, F), (u, U)\} \\ EX^f \text{ if } (x, x_r) = (f, U) \\ EX^u \text{ if } (x, x_r) = (u, F) \end{cases}.$$

These two theorems follow, the proofs of which may be found in Appendix 1.

THEOREM 5.1: In the fair luck game, agents in the second stage sanction, i.e., they allocate

[15] The assumption that $b < 1$ is not critical, but it makes the theoretical predictions consistent with the fact that, in such experiments, some second stage allocators set their goal above equality but, on average, somewhat below the maximum possible reward Z.

beyond what is called for by distributive preferences alone. Specifically, they reward first stage dictators more, or punish them less, for choosing f than for choosing u, ceteris paribus, i.e., for a given scale, x'.

THEOREM 5.2: Choices are sanctioned, even when the intended outcomes do not obtain, but fair choices are rewarded more strongly and unfair choices punished more strongly when realized outcomes are aligned with choices. Sanctions for realized outcomes that agree with expected outcomes are the same as for the same outcomes chosen with certainty.

These theorems predict most (and contradict none) of the stylized facts above as well as the specific findings of Cushman et al. while adding a theoretical underpinning for them in terms of fairness preferences. The theory is also consistent with the asymmetry in sanctioning in SF 5.1 from the concavity of *r*, but this cannot be proven for the fair luck game, given that it produces only a binary signal of preferences. Further corroborative evidence of this specification of virtue preferences is presented in section 6 on willful ignorance and section 7 on delegation.

6. Willful Ignorance

An important factor contributing to the 2007-08 financial crisis was the ability of lenders to avoid documenting applicants' incomes and, thereby, avoid knowing about the borrowers' often inflated claims. In the accounting sandals of the early 2000s, CEOs of troubled firms like Enron and Worldcom later professed ignorance of the dubious accounting practices at their companies. As these examples illustrate, the economic consequences of information avoidance can be staggering, but this phenomenon also spans other important domains. Political polarization, for example, can be traced to the shunning of broadcast and online news sources that uncomfortably challenge one's preconceptions (Dahlgren, Shehata, and Strömbäck, 2019, Peterson, Goel, and Iyengar, 2019). People often avoid information that could help them better identify the moral, right or socially beneficial course of action, which I will call *willful ignorance*. Experiments indicate that this behavior has an intrinsic component that cannot be explained, at least not solely, by extrinsic motives such as unadulterated greed, fear of legal culpability, or information costs.

Willful ignorance is one example of a broader class of anomalies I call "norm avoidance." Others include delegation, which is discussed in the following section, and moral egress, which is analyzed in Konow (2022). Norm avoidance denotes avoiding either making a

choice involving a salient moral norm or information about the consequences of such a choice. In the context of dictator experiments on this topic, I make the following two assumptions.

ASSUMPTION 1: Compared to moral salience in the standard dictator game (σ^h), moral salience is lower with the availability of an option to avoid taking action or acquiring information about the consequences of one's action (σ^m), even if the agent does not exercise the option. Moral salience is lower still for those who actually exercise the option and choose to avoid the action or information about the consequences of the action (σ^l).

That is, the option to avoid a moral norm, or information about its consequences, is non-moral context, which lowers moral salience. Actually exercising that option additionally increases separation between agent and patient and, therefore, lowers moral salience further.

ASSUMPTION 2: With norm avoidance, the effects in utility terms of different choices on altruism are second order in magnitude to the utility effects on material utility and for all choices are non-negative (non-positive) for altruistic (spiteful) dictators.

The examples of norm avoidance we consider involve categorical choices lacking clear endowments, so there are no unambiguous harming options, thus, the agent's altruism utility is assumed to be non-negative in all patient payoffs for $\alpha > 0$ and non-positive for $\alpha < 0$. Further, I assume in games with norm avoidance options that the effects of those options operate chiefly through fairness and material utility, and the latter swamps the effects of different choices on the utility from altruism.

Turning now specifically to willful ignorance, consider the example of the binary dictator game introduced by Dana, Weber and Kuang (2007), or DWK, which I will call the "information game." Most subsequent studies of willful ignorance in experimental economics employ this design or ones very close to it, although quite different designs, e.g., Serra-Garcia and Szech (2019) and Spiekermann and Weiss (2016), have also come to similar conclusions. There are three possible payoffs, H, F and L, that can be paired between D,R in four ways, $\{\pi_D, \pi_R\}$, where $0 \leq L < F < H$ and $F - L > H - F$ ($H = 6$, $F = 5$, $L = 1$ in DWK). The sequence of decisions and payoffs are illustrated in Figure 5. There are two states of the world, $\omega \in \{1,2\}$, that Nature (N) chooses with equal probability, $q = 0.5$. The D first chooses whether to reveal the realization of the gamble (R) or not (NR) and then chooses either option A or option B. If D chooses reveal, D then finds out whether the fairer option is A or B before choosing: in state 1, this is 1B, and, in state 2, this is 2A. As usual in a simple dictator game, fairness reduces to equality. Specifically in

the information game, I assume the entitlement is always the patient's payoff in the most equal state of the overall game, which is F in this case, i.e., $\eta = F$. This seems reasonable, it fits the results well, and alternative assumptions (such as equality in realized payoffs) complicate the analysis to the point that few choices can be ranked. If D chooses not to reveal, option A or B is chosen without knowledge of the realization of the gamble, whereby the fairer option is B in expectations. Since revealing is costless, though, D can always guarantee the fairer outcome, although that would mean sacrificing a payoff of H should state 1 obtain. For comparison, some studies include baseline treatments in which Ds know payoffs: the more common one relates to state 1 and D chooses only between R1A and R1B, call them 1A and 1B, respectively; in another the choices pertain to state 2 and are between R2A and R2B, call them 2A and 2B, respectively.

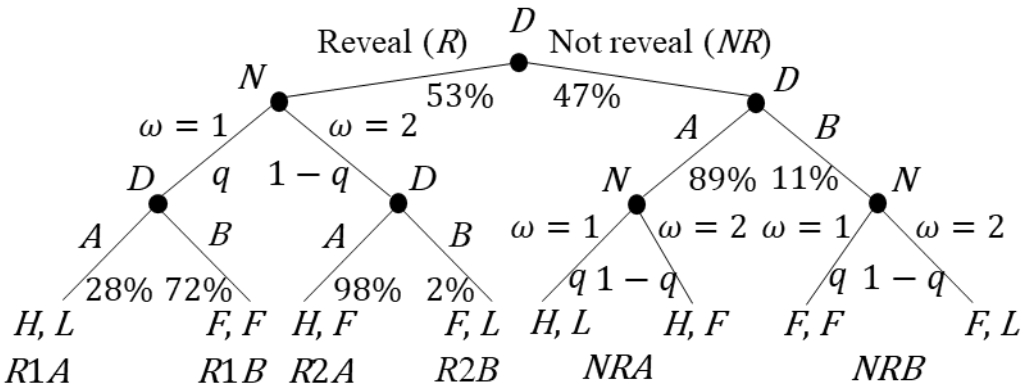

FIGURE 5. – Information game of Dana, Weber and Kuang (2007).

Consider the following stylized facts for this game. The percentages cited are averages weighted by numbers of observations (total N=368) from six studies: DWK plus five others that employ their design with relative payoffs that are the same or very close: Bartling, Engl and Weber (2014), Feiler (2014, $q = 0.5$, payoff sets 1 and 4), Grossman (2014, Default NR condition), Grossman and van der Weele (2017), and Larson and Capra (2009). These percentages as well as the acronyms for the various choices are summarized in Figure 5.

SF 6.1: In the information game, Ds are roughly equally split between those who reveal (53%) and those who do not (47%). Of those who reveal, a majority (72%) chooses the fair option B, if state 1 obtains (R1B), and nearly all (98%) choose option A, if state 2 obtains (R2A). Of those who do not reveal, a large majority (89%) chooses option A (NRA). In the baseline condition for state 1, most (69%) choose the fair option B (i.e., 1B), and, in the one study with a baseline for state 2 (Bartling et al., 2014), almost all (96%) choose option A (2A).

Let the D's choice in state ω be denoted x_ω, $i = \{1,2\}$, where the possible choices in each state are A or B. Then the D's utility when the state is revealed can be written

$$U(x_\omega, \omega) = u(\pi_D | x_\omega, \omega) + \sigma\phi f(\pi_R - \eta | x_\omega, \omega) + \sigma g(\pi_R, \alpha | x_\omega, \omega).$$

The D's expected utility before knowing the state can be written

$$EU(x_\omega, \omega) = 0.5 \cdot [u(\pi_D | x_1, 1) + u(\pi_D | x_2, 2)] + 0.5 \cdot \sigma[\phi f(\pi_R - \eta | x_1, 1) + \phi f(\pi_R - \eta | x_2, 2) + 0.5 \cdot \sigma[g(\pi_R, \alpha | x_1, 1) + g(\pi_R, \alpha | x_2, 2)].$$

Applied to willful ignorance, Assumption 1 means that the option in the information game to remain ignorant of the consequences of one's choices lowers moral salience, even for those who reveal (σ^m), compared to the standard game (σ^h), and moral salience is even lower for those who do not reveal (σ^l).

Unlike the previously discussed reciprocity experiments and sinners and saints game, which produce quasi-continuous measures of generosity, the information game does not produce any cardinal measures of generosity. The information game permits only categorical choices (whether or not to reveal, option A or B in state 1 or 2), but, unlike the fair luck game on outcome bias, the choices number more than two, and it is not immediately clear which choices should be considered more or less generous. Nevertheless, one can partially order choices according to their estimated generosity from inferences about intrinsic moral motivation using features of the theory discussed in section 2. As shown in the Theorem 6.1 below, the primary rankings here are according to the fairness coefficient. Experimental evidence on the resulting rankings comes later from third party sanctioning of choices. Moreover, coupled with Theorem 6.2 below, that evidence is shown to be highly consistent with both the predicted ranking of choices and the theory of virtue preferences.

The following theorem explains those findings in SF 6.1 involving (near) unanimity and provides a partial ordering of choices. The proof is located in Appendix 1.

Theorem 6.1: In the information game, if Ds do not reveal, option A (NRA) dominates (is strictly preferred to) option B (NRB). If Ds reveal and state 2 obtains, option A (R2A) dominates option B (R2B); R2B might only be chosen if there are very spiteful Ds. The fairest Ds reveal and choose the fairer option B if state 1 obtains (R1B), and the least fair Ds reveal and choose the less fair option A if state 1 obtains (R1A). In the baseline condition for state 1, fairer Ds choose B (1B) over A (1A). In the baseline condition for state 2, A (2A) dominates B (2B); 2B might only be chosen if there are very spiteful Ds. The estimated

generosity of choices C, $\hat{\gamma}^C$, can be ranked for dominant and dominated choices:

For dominant choices: $\hat{\gamma}^{R1B} > \hat{\gamma}^{NRA} > \hat{\gamma}^{R1A}$; $\hat{\gamma}^{R1B} > \hat{\gamma}^{R2A} > \hat{\gamma}^{R1A}$; $\hat{\gamma}^{1B} > \hat{\gamma}^{1A}$;

For dominated choices: $\hat{\gamma}^{R2A} > \hat{\gamma}^{R2B}$; $\hat{\gamma}^{2A} > \hat{\gamma}^{2B}$.

Proof: See Appendix 1.

Two choices, R2B and 2B, are predicted to be dominated, which results from Assumption 2. The fact that they are almost never chosen (only 2% and 4%, respectively) justifies this assumption, but these choices can, nevertheless, be ranked, because they would be chosen, if Ds were more spiteful than they, in practice, are. Nevertheless, one of the ten choices in the experiment, viz., NRB, cannot be ranked, because it is predicted always to be dominated. Although a large majority of Ds chooses NRA, roughly one in ten chooses NRB. Corollary 6.1 offers a means to rank this choice that is also consistent with a fraction of Ds choosing NRA.

Corollary 6.1: If some Ds in the information game value equality in final payoffs, then among them fairer Ds choose NRB, i.e., $\hat{\gamma}^{NRB} > \hat{\gamma}^{NRA}$.

Proof: See Appendix 1.

As explained previously, the analysis otherwise rests on the assumption that Ds associate fairness with the most equal payoff in the overall game, an assumption that enables a more extensive set of rankings of generosity and comports well with the results. But allowing for a few Ds, who value equal final payoffs, permits us to rank NRB and is consistent with its infrequent choice.

Bartling, Engl and Weber (2014, or BEW henceforth) adapted the DWK design, adjusting the payoffs while maintaining the fundamental relative relationships between L, F and H. They also added third party punishment of the D using the strategy method and based on D choices to reveal or not and of option A or B as well as of the realized payoffs. In their design, the possibility of punishment was common knowledge, and third parties paid a small price to punish. Below I analyze a simplified non-strategic version of third-party punishment in the information game, e.g., punishment by unannounced spectators. This makes the analysis more tractable while producing conclusions that should be qualitatively the same in light of several facts: actual punishment in BEW never changes the ranking of expected payoffs across fair and unfair choices relative to pre-punishment payoffs, the price of punishment in BEW is only \$0.10 per unit, and the third parties punish consistently with the ranking of choices from the non-strategic analysis here.

The following theorem states predictions about spectator punishment in this game.

Theorem 6.2: In the information game, suppose that F, which is the maximum threshold for generosity of any Ds, is the $\tilde{\gamma}$ for a fraction of spectators. Then, the ideal sanction of Ds by spectators, $\tilde{z}^{C\omega}$, varies according to D choice, C, and, in the case of NRA and NRB, realized state, $\omega \in \{1,2\}$, and can be ranked across choices and states as follows:

$$0 = \tilde{z}^{R1B} > \tilde{z}^{NRA2} > \tilde{z}^{NRA1} > \tilde{z}^{R1A}$$

$$0 = \tilde{z}^{R1B} > \tilde{z}^{R2A} > \tilde{z}^{R1A}$$

$$\tilde{z}^{R2A} > \tilde{z}^{R2B}$$

$$\tilde{z}^{NRB1} > \tilde{z}^{NRA2} > \tilde{z}^{NRA1}$$

$$\tilde{z}^{NRB1} > \tilde{z}^{NRB2} > \tilde{z}^{NRA1}$$

$$0 = \tilde{z}^{1B} > \tilde{z}^{1A}$$

$$\tilde{z}^{2A} > \tilde{z}^{2B}$$

Proof: See Appendix 1.

The assumptions about $\tilde{\gamma}$ allow one to conclude there will be no punishment of R1B and 1B but some punishment, if only minor, of other choices.

Table 6 presents the results from BEW on sanctions by D choice and, in the Not reveal cases, realization of state. Punishment is expressed in negative terms as punishment points and in order of increasing punishment from left to right. The top rubric summarizes the results for the information game and the bottom rubric the results for the baseline treatments. The results are consistent with all theoretical predictions. Punishment in R1B and 1B is effectively zero, and, although theory cannot produce a complete ordering across all twelve cases, the levels of punishment are consistent with all predicted partial orderings.

R1B	R2A	NRB1	NRB2	NRA2	R2B	NRA1	R1A
-0.58	-2.67	-4.42	-6.00	-8.00	-9.50	-11.42	-16.25

1B	2A	2B	1A
-0.56	-1.76	-12.41	-19.72

TABLE 6
PUNISHMENT POINTS IN BARTLING, ENGL AND WEBER (2014)

Numerous explanations have been offered for information avoidance, e.g., see the excellent review of Golman, Hagmann and Loewenstein (2017). On the more specific topic of willful ignorance, which as used here involves a connection to moral preferences, Gino, Norton

and Weber (2016) explain willful ignorance based on motivated reasoning, i.e., ignorance allows selfish dictators to believe they are being moral. This seems consistent with other evidence of self-serving fairness biases, e.g., Babcock, Loewenstein, Issacharoff, and Camerer (1995), Konow (2000), although I am unaware of any evidence on the incentivized elicitation of moral beliefs in the specific case of willful ignorance. Along other lines, Grossman and van der Weele (2017) propose a theory of self-image that is consistent with D behavior in the information game. They argue persuasively in favor of an explanation based on self-signaling, although it is unclear how that theory might explain patterns of third-party punishment. As I see it, though, the main arguments for moral salience are its parsimony and broad range of applications, while being potentially complementary to, rather than conflicting with, alternative explanations such as motivated reasoning and self-image.

7. Delegation

Numerous management consulting firms exist largely to recommend or carry out the firing of the employees of their client companies, even though those companies could implement the firings themselves and, thereby, save themselves the consulting fees. Companies in developed nations outsource much of their manufacturing to companies in less developed countries where labor standards are lower, even though there might, in some cases, be cost advantages from vertically integrating foreign production. When decision-makers delegate such choices, it raises the question of whether they seek to deflect blame from themselves for undesirable consequences, say, from their personal involvement in firings or from dangerous work conditions, such as those that led to the collapse of the Rana Plaza textiles factory building in Bangladesh in 2013 that killed more than 1100 workers. In fact, economics experiments corroborate the desire of agents to delegate immoral choices to others after ruling out other reasons, including liability concerns, the value of outside expertise and advantages of specialization.

Experimenters have studied delegation chiefly using dictator games in which a dictator may delegate to an intermediary (I) the decision about the payoffs of the D, R(s), and I. There have been wide variations, though, in features of the designs, such as continuous or binary choices, single shot or multiple rounds, communication between subjects or not, differing numbers of subjects in groups, fixed matching or rematching of groups, selection of Is, different

opportunities for punishment and of different members of groups, etc. Nevertheless, certain patterns are robust across these designs and are summarized in the following stylized facts.

SF 7.1: When dictators have an option to delegate, average allocations to recipients are lower than in a standard dictator game. Some dictators delegate the allocation decision to intermediaries, who usually choose unfair allocations, and fewer dictators make fair allocations directly themselves (Hamman, Loewenstein and Weber, 2010, Coffman, 2011, Bartling and Fischbacher, 2012, Oexl and Grossman, 2013).

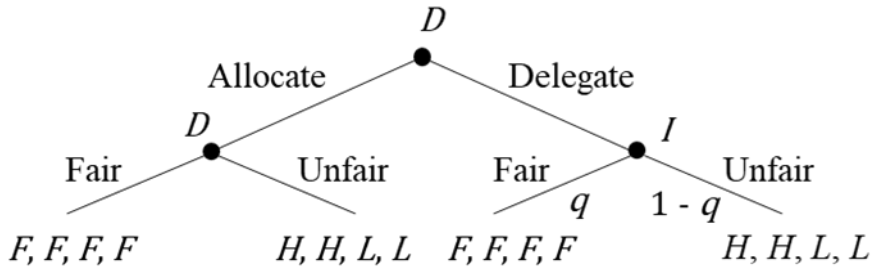

FIGURE 6. – Delegation game of Bartling and Fischbacher (2012).

For the analysis, I focus on a simple, non-strategic design, which I will call the "delegation game," that was introduced by Bartling and Fischbacher (2012, henceforth BF) and that has been used and adapted by others, e.g., Oexl and Grossman (2013). Each group of four consists of a D, an I, and two Rs, whose payoffs, $\{\pi_D, \pi_I, \pi_{R1}, \pi_{R2}\}$, can be fair, $\{F, F, F, F\}$, or unfair, $\{H, H, L, L\}$, where $0 \leq L < F < H$ and $L + H = 2F$ ($H = 9$, $F = 5$, $L = 1$ in BF). The sequence of decisions and payoffs are illustrated in Figure 6. The D chooses either to allocate directly or to delegate the decision to an I, whereby the D is assumed to estimate the probability that the I will allocate fairly to be q, where $0 < q < 1$. There is also a baseline treatment with no delegation option that corresponds to the Allocate branch of the game tree. Clearly, standard theory predicts that a risk neutral or risk averse D will never delegate, since allocating directly guarantees the fair or unfair outcome preferred by D. BF include treatments with punishment that produce high rates of fair choices, likely motivated by strategic self-interest. But in their non-strategic treatments of the delegation game without punishment, 66% of Ds choose unfair, 17% choose fair, and 17% choose to delegate, whereby 82% of Is then choose the unfair allocation. In the baseline dictator game, 65% of Ds choose unfair, almost identical to that in the delegation game, but that means that the percentage of Ds choosing fair directly drops from 35% in the baseline to only 17% in the delegation game. Thus, the delegation option appears to lead about

one-half of otherwise fair Ds to delegate.

Applied to the delegation game, Assumption 1 means that the very existence of an option to delegate the decision in the delegation game to an intermediary lowers moral salience (σ^m) relative to the standard dictator game (σ^h), and moral salience is lower still for those who actually choose to delegate (σ^l). D's uncertainty about I's choice might magnify the impact of intermediation through the previously discussed moral uncertainty, but delegation alone has been shown to affect behavior in the predicted manner even in the absence of uncertainty (e.g., see results in Study 1 of Coffman, 2011).[16] The moral salience experienced by the intermediary, σ^I, should lie within a range: since I is aware of the delegation option, salience should be no greater than σ^m, but, the actual selection of the delegation option might reduce the sense of moral responsibility of both D and I by a common degree to σ^l. Hence, I assume that $\sigma^I \in [\sigma^l, \sigma^m]$. Since choices are categorical and not related in obvious ways to generosity, I proceed as with willful ignorance by inferring the generosity of different choices theoretically based on intrinsic moral preferences. In the delegation game, choices can be ranked based on fairness preferences alone to produce estimated generosity, $\hat{\gamma}_P^{GC}$, where *G* equals the baseline (B), allocation by the D in the delegation game (A), or delegation by the D in the delegation game (D), *C* represents fair (F) or unfair (U) in the cases where the subject makes that choice directly (i.e., omitted in the case of the Ds who delegate), and P indicates that the estimated generosity refers to that of the dictator (D) or intermediary (I).

Theorem 7.1: In the delegation game, the fairest dictators choose to allocate fairly themselves, less fair ones delegate, and the least fair allocate unfairly themselves. Fewer dictators choose to allocate fairly in the delegation game than in the standard dictator game. The fraction of intermediaries allocating fairly is greater than or equal to the fraction of Ds choosing to allocate fairly in the delegation game and strictly less than the fraction of fair Ds in the baseline. Estimated generosity can be ranked as follows:

For dictators: $\hat{\gamma}_D^{BF} > \hat{\gamma}_D^{AF} > \hat{\gamma}_D^{D} > \hat{\gamma}_D^{AU}; \hat{\gamma}_D^{BF} > \hat{\gamma}_D^{BU} > \hat{\gamma}_D^{AU}$

For intermediaries: $\hat{\gamma}_D^{DF} \geq \hat{\gamma}_I^{DF} > \hat{\gamma}_D^{D}; \hat{\gamma}_I^{DF} \geq \hat{\gamma}_I^{DU} > \hat{\gamma}_D^{AU}$.

[16] One question is whether to treat I as an agent or a patient. In the latter, but not the former, case, I's allocations must be included in the moral preferences of the D. In the baseline, it seems clear that I is wholly passive and, therefore, a patient. The same is true when the D chooses directly in the delegation game. For simplicity and consistency, therefore, and because the results do not depend qualitatively on this call, I is treated everywhere as a patient.

Proof: See Appendix 1.

These predictions are consistent with the patterns observed in the BF treatments discussed above. Some Ds allocate fairly in the baseline (35%). When delegation is an option in the delegation game, some delegate and the fraction of fair Ds falls (17%) to a level below the fraction of fair Ds in the baseline and roughly equal to the fraction of fair Is (18%), suggesting σ^I is close to σ^l.

Many experimental studies of delegation also include the possibility of subsequent punishment of dictator decisions. Some of these results are summarized in the following SF.

SF 7.2: In the delegation game, there is no significant punishment of fair choices, regardless of delegation. Dictators are punished significantly less for unfair allocations that result from delegation than from their own decisions, but delegation increases punishment of intermediaries for unfair choices (Coffman, 2011, Bartling and Fischbacher, 2012, Oexl and Grossman, 2013).

The following theorem considers costless punishment in the delegation game assuming non-strategic D choices as with unanticipated spectator punishment.

Theorem 7.2: In the delegation game, suppose that F, which is the maximum threshold for generosity of any Ds, is the $\tilde{\gamma}$ for at least some spectators. Then, spectators do not punish those who choose fair directly or those who do not choose, but, depending on their threshold for generosity, $\tilde{\gamma}$, might punish some choices in the five remaining cases, which can be ranked as follows:

$$\tilde{z}_D^{DF} > \tilde{z}_D^{DU} > \tilde{z}_D^{AU}, \tilde{z}_D^{DU} > \tilde{z}_D^{BU}, \tilde{z}_I^{DU} > \tilde{z}_D^{AU}.$$

Proof: See Appendix 1.

These imply, among other things, stronger punishment of unfair allocations that are directly chosen in the baseline or in the delegation game than those resulting from delegation. Since the threshold for generosity is an empirical question, certain unfair choices might not be punished, such as a D's choosing to delegate, which is the most generous unfair choice.

I am unaware of any delegation experiment with unannounced spectator punishment, but in some treatments of BF, a randomly chosen R could pay a fixed fee of one point to assign up to seven punishment points to each of the other three subjects for each possible decision (i.e., using the strategy method). In these treatments, punishment was common knowledge, which potentially confounds inferences about the motives behind D choices because of possible strategic self-interest. Nevertheless, if R estimates of D types in these treatments produces the

same ranking of Ds as that of unannounced spectators, then the results remain qualitatively the same. Thus, this comparison must be taken with a grain of salt, but these results are worth examining given the negligible cost of punishment in BF, the absence of a compelling reason to believe the possibility of strategic self-interest would undo rankings of Ds, and the consistency of predictions with the patterns observed in this study. The results for punishment in the BF experiment are summarized in Table 7 and are consistent with the predictions of Theorem 7.2. Punishment is negligible when choices are fair or when subjects do not choose, and direct choices of unfair, whether by Ds or Is, are punished more strongly than delegated ones. As predicted, when the D delegates, the D's punishment is greater, if the I subsequently chooses unfair than fair, since the scale of the former is F and that of the latter only EX^u, indeed, $\tilde{z}_D^{DF}$ is negligible. But the overall level of punishment of the D with delegation is very modest, suggesting the spectator threshold for generosity is such that they do not, on average, consider delegation a very unfair choice.

	Fair		Unfair	
	Dictator	Intermed	Dictator	Intermed
Dictator Game:	BF/D	BF/I	BU/D	BU/I
Baseline	-0.41	-0.34	-3.70	-0.42
Delegation Game:	AF/D	AF/I	AU/D	AU/I
Allocate	-0.19	-0.15	-4.27	-0.75
Delegation Game:	DF/D	DF/I	DU/D	DU/I
Delegate	-0.24	-0.20	-1.31	-3.96

TABLE 7
PUNISHMENT POINTS IN BARTLING AND FISCHBACHER (2012)

8. Moral Point Salience

So far, moral salience has referred in this paper to what is more properly called *moral set salience*, but now we turn briefly to what I will call *moral point salience*. As I use the terms here, set salience relates to properties of disjoint subsets of the decision context, viz., moral and non-moral context, whereas point salience refers to individual elements of the context. The latter provides a simple explanation for the well-established pattern from economics experiments of

atoms at certain points in choice distributions. I will discuss three examples of actions that are chosen with greater frequency than alternative choices in their neighborhood. I believe these are important examples of point salience, but I do not claim that this is necessarily an exhaustive list. First, equal splits are a frequent choice in many experiments, such as in ultimatum and dictator games (Camerer, 2003). Second, zero transfers have also emerged as a frequent choice in dictator experiments where taking options rule out a corner solution as the sole explanation, e.g., List (2007), Cappelen et al. (2013b), and Alevy et al. (2014). Finally, many studies have found that, when certain actions are explicitly highlighted (e.g., actions of previous subjects, experimenter suggestions, actions of role models), they tend to be chosen more frequently, including in dictator games, e.g., Andreoni and Bernheim (2009), public good games, e.g., Croson and Marks (2001), COVID-19-related contributions and volunteering, e.g., Abel and Brown (2020), and field experiments on charitable contributions, e.g., Shang and Croson (2009).

Researchers have explained or modeled these patterns in various ways, but each account has its limitations, and I am unaware of a unified explanation for all three. For example, theoretical explanations for equal splits include a kinked inequality aversion term (Fehr and Schmidt, 1999), infinite inequity aversion on the part of some subjects (e.g., Konow, 2000), or a signaling game in which agents value social image (Andreoni and Bernheim, 2009). A strength of the first approach is that it provides a simple explanation for results from numerous bargaining and market experiments, but it is inconsistent with other findings, such as the frequent choice of transfers between zero and one-half in dictator experiments. The second approach accounts for the heterogeneity of types according to their degree of inequity aversion observed in many experiments and also accommodates concepts of equity other than equality, but it is inconsistent with variation in the percentage of subjects making equitable choices, for example, with the price of giving (Andreoni and Miller, 2002). The third approach offers a persuasive account for equal splits in the dictator game that avoids seemingly ad hoc assumptions about non-differentiability of utility, but it relies on a complicated theoretical apparatus the extension of which to other games is not straightforward. Moreover, none of these approaches provides explanations for all three examples above. Zero transfers, even in dictator games with taking, might be explained by dictators, who experience an endowment effect, but that explanation does nothing to account for the first or third examples. The masses at highlighted choices can be understood as focal points that facilitate coordination in strategic games, but that does not explain masses at dominated

choices in non-strategic decisions. One might invoke experimenter demand effects (Zizzo, 2010), but they also do not provide a coherent and unified explanation for all three effects.

Moral point salience offers a simple and parsimonious account for these three types of masses based on salience and moral norms. Specifically, it concerns elements, P, of the set of actions, X, that are, for moral reasons, more salient than other elements of X, whereby $P \subset X$. Moral point salience is a term applied to fairness preferences, f, of an agent i and takes the form:

$$s_i(x) = \begin{cases} \overline{s}_i \geq 1 \\ 1 \end{cases} \text{if} \begin{matrix} x \in P \\ x \notin P \end{matrix}.$$

That is, morally salient actions may be more heavily weighted in some agents' fairness preferences than other elements of the set of actions, whereby the weight can vary by person.

This discontinuity in utility can prove irksome, and, in fact, the analysis in this paper thus far has not required point salience to be invoked, largely for that reason. Without point salience, the theory has been applied to explain a wide range of classic and anomalous results while retaining differentiability of the utility function. Nevertheless, point salience is helpful to account for masses that often materialize when examining the distribution of choices. So, I wish to outline and justify briefly the position that moral point salience not only earns first place in an Occam's razor contest to explain such masses but also that it is also a persuasive part of a coherent morals-based framework. I propose three categories of moral point salience below.

First, *norm salience*, where $\eta \in P$, is the most intuitive type of moral point salience. When first hearing about the standard dictator or ultimatum game, I suspect almost everyone thinks the same thing: the morally right choice is to split the stakes equally. We can torture ourselves for alternative, and more elegant, explanations, but I believe the most compelling one is staring us in the face: the morally preferred choice is obvious, in this case. I take obvious to imply, formally, that there is a discrete decrement in utility for making another choice, or, equivalently, a discrete increment for making the morally obvious choice. Of course, stakeholders, such as dictators or proposers, might make another choice due to self-interest, but, as stated in Assumption 3, the moral norm, η, can be identified from the choices of spectators. A concrete and intuitive way to operationalize this is to associate the entitlement in experiments with the modal choice of spectators and the salience of the entitlement as being in direct proportion to spectator consensus, specifically, whereby consensus can be conceptualized as inversely related to variance in spectator judgments as proposed in Konow (2009).

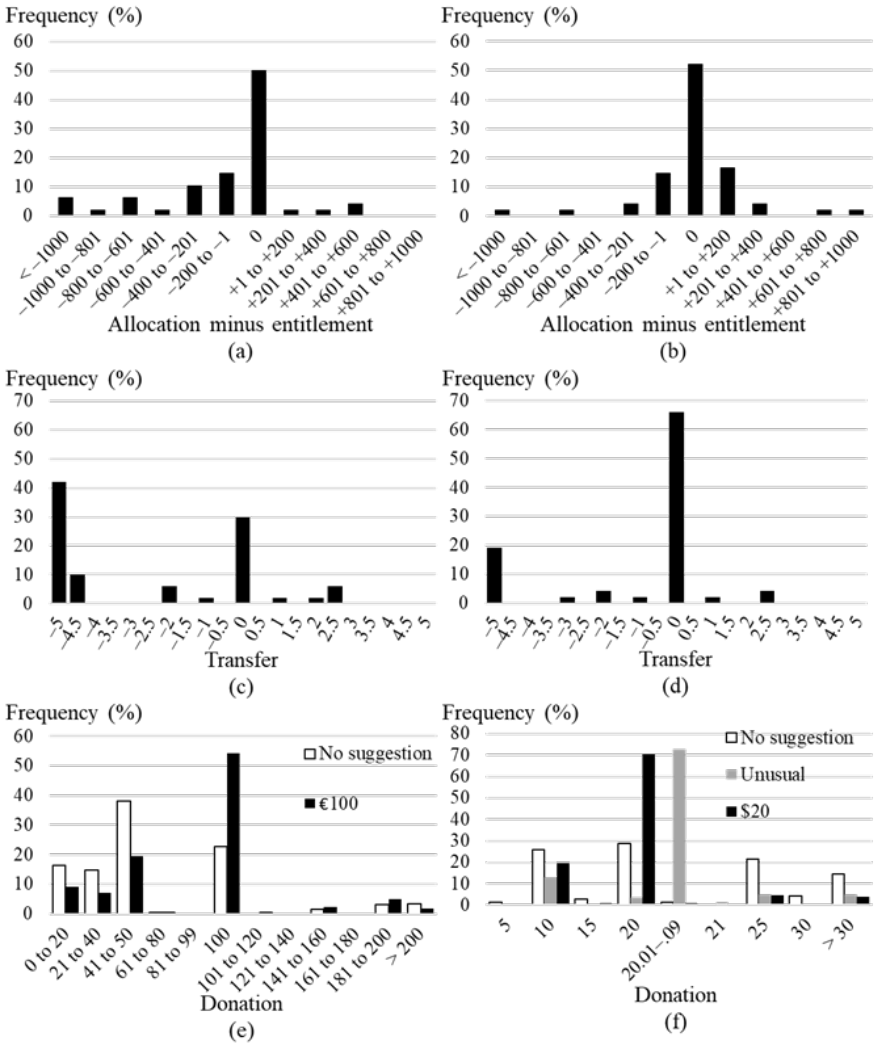

FIGURE 7. – Moral point salience.

Sources: Konow. Saijo and Akai (2020) stakeholders (a) and spectators (b), List (2007) Take $5 (c) and Earnings (d), Adena, Huck and Rasul (2014) No suggestion and €100 (e), and Edwards and List (2014) No ask, $20 ask and Unusual ask (f).

As previously discussed, the norm defaults to equality in simple decision contexts, like the standard dictator game. In fact, equal splits emerge frequently in most of the games examined in this paper. But what if the norm is not as simple and obvious as equality? As argued in section 3, when the context provides information relevant to other norms, behavior shifts towards those norms, but does that produce masses at those norms? Consider a more demanding test of norm salience based on a more complicated rule: equity calls for allocations that are proportional to contributions that differ across agents. Figure 7 summarizes results from experiments that illustrate this rule. In Konow, Saijo and Akai (2020), subjects first perform a real effort task, and then dictators allocate the resulting earnings. Panel a shows the amounts stakeholding dictators

allocate to recipients, and panel b shows the amounts spectators allocate to one member of each pair. Specifically, the horizontal axis represents the difference in points allocated to recipients from the amount they produced, which is their entitlement and the equitable amount. The mode and choice of 50% of stakeholders in panel a and of 52% of spectators in panel b is the equitable amount. Thus, in general, with norm salience, P is a rule and x a point that might be conditioned on another variable in the context. For example, when the norm is equity, x is conditioned on individual contributions, and when the norm is basic needs, it is conditioned on individual needs. Konow (2001) and Konow, Saijo and Akai (2020) argue that equality is the norm by default, when there is no or insufficient information to apply principles that depart from equality.

Second, *null salience* is the salience of inaction, i.e., the choice neither to help nor to harm, such as neither giving nor taking in a dictator game, denoted $0 \in P$. Null salience is related to the distinction in ethics between sins of commission for the wrongs one chooses versus sins of omission for the acts one should perform but does not. Various experiments in economics and psychology suggest people have a stronger aversion to acts of commission than to ones of omission (e.g., Cox, Servátka, and Vadovič, 2017, Spranca, Minsk, and Baron, 1991). That is, an individual, who otherwise might prefer to harm another, say, take $1 in a dictator game, might experience a discontinuity in the marginal moral disutility of doing so. One way to model this is with moral point salience, where utility is discretely greater at zero. Of course, in many experiments, such as the standard dictator game, a mass at zero logically emerges as a corner solution due to selfish, spiteful or insufficiently fair dictators. But such censoring is not a problem in dictator games with taking, and, in fact, those studies typically also find a mass at zero. This is illustrated in panels c and d of Figure 7, which depict the Take $5 and Earnings treatments, respectively, of List (2007), and show between 30% and 66% of dictators choosing inaction, i.e., transfers of zero.[17]

Third, and finally, *threshold salience* refers to the action, $\tilde{x} \in P$, that corresponds to the agent's preferred character threshold, $\tilde{\gamma}$, given the context and its moral set salience. Remember that $\tilde{x}$ is the action that is neither praiseworthy nor blameworthy and that it is less than the objectively fair transfer, η, if salience is below a sufficiently high level. Thus, this can be thought

[17] Another way this could be modeled is as a kink in the altruism function at zero. This is consistent with a mass at zero but one disadvantage of this approach is that a kink is not consistent with the paucity of transfers typically observed just above and just below zero whereas point salience is.

of as the action that is "fair enough" given the context. In the case of a stakeholder, we can think of this as the stakeholder's (potentially biased) belief about what constitutes "fair enough." The current theory assumes heterogeneity in the value of $\tilde{\gamma}$, even among spectators, so it is not clear why a mass would materialize at any particular stakeholder choice. But experimental evidence establishes that beliefs about the sufficient level of norm compliance, $\tilde{x}$, are malleable and can be manipulated, e.g., Bicchieri and Chavez (2010) and Bicchieri, Dimant, Gächter and Nosenzo (2020). Indeed, compliance with norms responds to and sometimes coalesces around information, including about past trusting behavior of others (Berg, Dickhaut and McCabe, 1995), recommended contributions to a public good (e.g., Croson and Marks, 2001) and default levels of transfers to recipients in dictator games (e.g., Andreoni and Bernheim, 2009). Such evidence is consistent with an effect of specific information on beliefs about appropriate norm compliance and, as a result, on behavior itself.

Voluntary contributions to charities or public goods lend themselves to examination of this effect, since many people feel morally obliged to donate but plausibly have non-degenerate distributions of beliefs about the appropriate amount. Studies of such contributions that include suggested donations yield evidence consistent with threshold salience. Panel e of Figure 7 shows contributions to a public good (viz., an opera house) in a field experiment of Adena, Huck and Rasul (2014). When solicitations explicitly suggest a €100 contribution, the fraction of such donations is significantly greater than that when no suggestion is made ($p<0.001$), according to the two-tail z-tests used in all comparisons here.[18] The results of the field experiment of Edwards and List (2014) on alumni donations to a university are summarized in panel f of Figure 7. The fraction of $20 contributions is significantly greater ($p<0.001$), when that amount is explicitly stated in solicitations. A further treatment in which solicitations suggest unusual amounts, like $20.01 or $20.04, produce a similar increase in the frequency of choices of stated amounts ($p<0.001$), corroborating the robustness of this effect, even when suggestions are not round numbers.

An advantage of the way these three types of moral point salience are formulated is that they can be specified and identified empirically. Norm salience can be inferred from spectator

[18] A further treatment shows that, when solicitations suggest €200, the effect dissipates. This seems consistent with agents having prior beliefs about the distribution of appropriate donations, whereby suggestions provide signals that impact $\overline{s}_i$ in direct relationship to their proximity to priors such that outlier suggestions are ineffectual. Nevertheless, formal analysis of this question goes beyond the scope of this paper.

choices, null salience defines itself, and threshold salience can be inferred from behavioral responses to information or from incentivized elicitation of beliefs. This discussion on moral point salience indicates some commonsensical concepts to account for masses that are not a part of the theory of moral set salience and indicates possible avenues for future research.

9. Conclusions

This paper proposes a tractable theory to explain not only classic results on allocative preferences and reciprocity but also a wide range of anomalous findings about moral behavior, including moral proximity, moral uncertainty, the outcome bias, the taking effect, joy of destruction, moral egress, willful ignorance and delegation. At various stages, I have discussed alternative explanations for specific phenomena, such as experimental artefacts (e.g., Bardsley, 2008), motivated reasoning (e.g., Gino et al., 2016) and image concerns (e.g., Andreoni and Bernheim, 2009), including what I see as the strengths of those alternatives. As stated at the start, the goal is not to dismiss or conduct a beauty contest with other accounts of specific phenomena. Instead, one goal was to present an until now neglected explanation, which plausibly sweeps up much of the variance in observed behavior. Another goal was to illustrate the theory's flexibility and ease of application, that is, to argue its appeal on the basis of Occam's razor. A related aim was to demonstrate the generality of the theory across an unprecedented set of sometimes enigmatic empirical results on moral preferences. Finally, the theory was tested out-of-sample and its predictions corroborated in a new experiment.

Future research could explore possible roles for moral salience and virtue preferences in relation to other types of moral preferences apart from allocative preferences, e.g., trust, trustworthiness, honesty, and cooperation. Further work could also analyze the factors that affect how different moral and non-moral contexts might be integrated across different decisions at a point in time as well as over time. That is, one could examine the effects on moral salience of presenting similar decisions while varying the moral and non-moral context, which could, for example, account for order effects. In addition, this paper focused on non-strategic decision-making in order to simplify the analysis and avoid factors that might confound inferences about the forces being studied. But future work might extend the theory to situations involving strategic interaction, such as bargaining. A theory incorporating moral salience and virtue preferences could be applied to decision-making in experimental games, like the ultimatum

game, trust game, moonlighting game, centipede game, and public good games.

REFERENCES

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner (2000): "The Moonlighting Game: An Experimental Study on Reciprocity and Redistribution," *Journal of Economic Behavior and Organization,* 42(2), 265-277.

Abbink, Klaus and Abdolkarim Sadrieh (2009): "The Pleasure of Being Nasty," *Economic Letters,* 105(3), 306-308.

Abbink, Klaus and Benedikt Herrmann (2009): "Pointless Vendettas," *SSRN Electronic Journal,* 1-11.

Abbink, Klaus and Benedikt Herrmann (2011): "The Moral Costs of Nastiness," *Economic Inquiry,* 49(2), 631-633.

Abel, Martin and Willa Brown (2020): "Prosocial Behavior in the Time of COVID-19: The Effect of Private and Public Role Models," *IZ Discussion Paper,* 13207, 1-26.

Adena, Maja, Steffen Huck, and Imran Rasul (2014): "Charitable Giving and Nonbinding Contribution-Level Suggestions — Evidence from a Field Experiment," *Review of Behavioral Economics,* 1(3), 275-293.

Aguiar, Fernando, Alice Becker, and Luis Miller (2013): "Whose Impartiality? An Experimental Study of Veiled Stakeholder, Involved Spectators and Detached Observers," *Economics and Philosophy,* 29 (2), 155-174.

Alevy, Jonathan E., Francis L. Jeffries, and Yonggang Lu (2014): "Gender – and Frame-Specific Audience Effects in Dictator Games," *Economic Letters,* 122, 50-54.

Almås, Ingvild, Alexander Cappelen, and Bertil Tungodden (2020): "Cutthroat Capitalism versus Cuddly Socialism: Are American More Meritocratic and Efficiency-seeking than Scandinavians?." *Journal of Political Economy,* 128 (5), 1753-1788.

Almenberg, Johan, Anna Dreber, Coren L. Apicella, and David G. Rand (2011): "Third Party Reward and Punishment: Group Size, Efficiency and Public Goods," *Psychology of Punishment,* 10, 1-17.

Andreoni, James (1989): "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy,* 97(6), 1447-1458.

Andreoni, James and B. Douglas Bernheim (2009): "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica,* 77(5), 1607-1636.

Andreoni, James and John Miller (2002): "Giving According to Garp: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica,* 70(2), 737-753.

Andreoni, James, Justin M. Rao, and Hannah Trachtman (2017): "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving," *Journal of Political Economy,* 125(3), 625-653.

Arechar, Antonio A., Simon Gächter, and Lucas Molleman (2018): "Conducting Interactive Experiments Online," *Experimental Economics,* 21, 99-131.

Ashraf, Nava, and Oriana Bandiera (2017): "Altruistic Capital," *American Economic Review: Papers & Proceedings,* 107(5), 70-75.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer (1995): "Biased Judgements of Fairness in Bargaining," *The American Economic Review,* 85(5), 1337-1343.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2010): "Social Incentives in the Workplace," *The Review of Economic Studies,* 77, 417-458.

Bardsley, Nicholas (2008): "Dictator Game Giving: Altruism or Artefact?," *Experimental Economics,* 11(2), 122-133.

Bartling, Björn and Urs Fischbacher (2012): "Shifting the Blame: One Delegation and Responsibility," *Review of Economic Studies,* 79, 67-87.

Bartling, Björn, Florian Engl, and Roberto A. Weber (2014): "Does Willful Ignorance Deflect Punishment? - An Experimental Study," *European Economic Review,* 70, 512-524.

Bénabou, Roland and Jean Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review,* 96(5), 1652-1678.

Benz, Mathias and Stephan Meier (2008): "Do People Behave in Experiments as in the Field? – Evidence from Donations," *Experimental Economics,* 11(3), 268-281.
Berg, Joyce, John Dickhaut, and Kevin McCabe (1995): "Trust, Reciprocity, and Social History," *Games and Economic Behavior,* 10, 122-142.
Bertrand, Marianne and Sendhil Mullainathan (2001): "Are CEOs Rewarded for Luck? The Ones Without Principals Are," *Quarterly Journal of Economics,* 116(3), 901-932.
Bicchieri, Cristina and Alex Chaves (2010): "Behaving as Expected: Public Information and Fairness Norms," *Journal of Behavioral Decision Making,* 23(2), 161-178.
Bicchieri, Cristina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo (2020): "Observability, Social Proximity, and the Erosion of Norm Compliance," *CESifo Working Paper,* 8212, 1-32.
Bohnet, Iris and Bruno S. Frey (1999): "Social Distance and Other-Regarding Behavior in Dictator Games: Comment," *The American Economic Review,* 89, 335-339.
Bolton, Gary E., Jordi Brandts, and Axel Ockenfels (2005): "Fair Procedures: Evidence from Games Involving Lotteries," *The Economic Journal,* 115(506), 1054-1076.
Bontol, Gary E., and Elena Katok (1998): "An Experimental Test of the Crowding Out Hypothesis: The Nature of Beneficent," *Journal of Economic Behavior & Organization,* 37(3), 315-331.
Bolle, Friedel, Johnathan H.W. Tan, and Daniel John Zizzo (2014): "Vendettas," *American Economic Journal: Microeconomics,* 6(2), 93-130.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Sheifer (2012): "Salience in Experimental Tests of the Endowment Effect," *American Economic Review,* 102(3), 47-52.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Sheifer (2013): "Salience and Consumer Choice," *Journal of Political Economics,* 121(5), 803-843.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Sheifer (2016): "Competition for Attention," *Review of Economic Studies,* 83(2), 481-513.
Brañas-Garza, Pablo (2007): "Promoting Helping Behavior with Framing in Dictator Games," *Journal of Economic Psychology,* 28, 477-486.
Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson (2007): "Is Generosity Involuntary?," *Economic Letters,* 94, 32-37.
Brock, J. Michelle, Andreas Lange, and Erkut Y. Ozbay (2013): "Dictating the Risk: Experimental Evidence on Giving in Risky Environments," *American Economic Review,* 103, 415-437.
Brownback, Andy and Michael A. Kuhn (2019): "Understanding Outcome Bias," *Games and Economic Behavior,* 117, 342-360.
Candelo, Natalie, Catherine Eckel, and Cathleen Johnson (2018): "Social Distance Matters in Dictator Games: Evidence from 11 Mexican Villages," *Games,* 9(77), 1-13.
Camerer, Colin (2003): *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton: Princeton University Press.
Camerer, Colin, and Richard H. Thaler (1995): "Anomalies: Ultimatums, Dictators and Manners," *Journal of Economic Perspectives,* 9(2), 209-219.
Campos-Mercade, Pol, Armando N. Meier, Florian H. Schneider, and Erik Wengström (2020): "Prosociality Predicts Health Behaviors During the COVID-19 Pandemic," Department of Economics Working Paper No. 346, University of Zurich.
Cappelen, Alexander W., James Konow, Erik O. Sorensen, and Bertil Tungodden (2013): "Just Luck: An Experimental Study of Risk-Taking and Fairness," *American Economic Review,* 103(4), 1398-1413.
Charness, Gary and David I. Levine (2007): "Intention and Stochastic Outcomes: An Experimental Study," *The Economic Journal,* 117(522), 1051-1072.
Charness, Gary and Martin Dufwenberg (2006): "Promises and Partnership," *Econometrica,* 74(6), 1579-1601.
Charness, Gary and Matthew Rabin (2002): "Understanding Social Preferences with Simple Tests," *The Quarterly Journal of Economics,* 117(3), 817-869.
Charness, Gary and Uri Gneezy (2008): "What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Games," *Journal of Economic Behavior and Organization,* 68, 29-35.

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016): "oTree – An Open-source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance,* 9, 88-97.
Chetty, Raj, Adam Looney, and Kory Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review,* 99(4), 1145-1177.
Cherry, Todd L., Peter Frykblom, and Jason F. Shogren (2002): "Hardnose the Dictator," *The American Economic Review,* 92(4), 1218-1221.
Chowdhury, Subhasish M., Joo Young Jeon, and Bibhas Saha (2017): "Gender Differences in the Giving and Taking of the Dictator Game," *Southern Economic Journal,* 84(2), 474-483.
Coffman, Lucas C. (2011): "Intermediation Reduces Punishment (and Reward)," *American Economic Journal: Microeconomics,* 3(4), 77-106.
Cox, James C. (2004): "How to Identify Trust and Reciprocity," *Games and Economic Behavior,* 46(2), 260-281.
Cox, James C., Maroš Servátka, and Radovan Vadovič (2017): "Status Quo Effects in Fairness Games: Reciprocal Responses to Acts of Commission Versus Acts of Omission," *Experimental Economics,* 20, 1-18.
Cox, James C., John A. List, Michael Price, Vjollca Sadiraj, and Anya Samek (2019): "Moral Costs and Rational Choice: Theory and Experimental Evidence," *Experimental Economics Center Working Paper Series,* 2, 1-52.
Crawford, Vincent P., and Nagore Iriberri (2007), "Fatal Attraction: Salience, Naïveté, and Sophistication in Experimental "Hide-and-Seek" Games," *American Economic Review,* 97(5), 1731-1750.
Crawford, Vincent P., Uri Gneezy, and Yuval Rottenstreich (2008): "The Power of Focial Points is Limited: Even Minute Payoff Asymmetry May Yield Coordination Failures," *American Economic Review,* 98(4), 1443-1448.
Croson, Rachel, and James Konow (2009): "Social Preferences and Moral Biases," *Journal of Economic Behavior and Organization*, 69(3), 201-212.
Croson, Rachel and Melanie Marks (2001): "The Effect of Recommend Contributions in the Voluntary Provision of Public Goods," *Economic Inquiry,* 39(2), 238-249.
Crumpler, Heidi and Philip J. Grossman (2008): "An Experimental Test of Warm Glow Giving," *Journal of Public Economics,* 92(5-6), 1011-1021.
Cushman, Fiery, Anna Dreber, Ying Wang, and Jay Costa (2009): "Accidental Outcomes Guide Punishment in a 'Trembling Hand' Game," *PLoS ONE,* 4(8), 1-7.
Dahlgren, Peter M., Adam Shehata, and Jesper Strömbäck (2019): "Reinforcing Spirals at Work? Mutual Influences between Selective New Exposure and Ideological Leaning," *European Journal of Communications,* 34(2), 159-174.
Dal Bó, Ernesto, and Pedro Dal Bó (2014): "Do the Right Thing: The Effects of Moral Suasion on Cooperation," *Journal of Public Economics,* 117, 28-38.
Dal Bó, Pedro, and Guillaume R. Fréchette (2018): "On the Determinants of Cooperation in Infinitely Repeated Games: A Survey," *Journal of Economic Literature,* 56(1), 60-114.
Dana, Jason, Daylian M. Cain, and Robyn M. Dawes (2006): "What You Don't Know Won't Hurt Me: Costly (But Quiet) Exit in Dictator Games," *Organizational Behavior and Human Decision Processes,* 100(2), 193-201.
Dana, Jasion, Roberto A. Weber, Jason Xi Kuang (2007): "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory,* 33, 67-80.
Dejean, Sylvain (2020): "The Role of Distance and Social Networks in the Geography of Crowfunding: Evidence from France," *Regional Studies,* 54(3) 329-339.
DellaVigna, Stefano, John A. List, and Ulrike Malmendier (2012): "Testing for Altruism and Social Pressure in Charitable Giving," *The Quarterly Journal of Economics,* 127, 1-56.
De Oliveria, Angela C.M., Alexander Smith, and John Spraggon (2017): "Reward the Lucky? An Experimental Investigation of the Impact of Agency and Luck on Bonuses," *Journal of Economic Psychology,* 62, 87-97.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth (2018): "Measuring and Bounding Experimenter Demand," *American Economic Review,* 108(11), 3266-3302.
Dreber, Anna, Tore Ellingsen, Magnus Johannesson, and David G. Rand (2013): "Do People Care About Social Context? Framing Effects in Dictator Games," *Experimental Economics,* 16(3), 349-371.
Dufwenberg, Martin and Georg Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior,* 47(2), 268-298.
Eckel, Catherine C. and Philip J. Grossman (1996): "The Relative Price of Fairness: Gender Differences in a Punishment Game," *Journal of Economic Behavior and Organization,* 30(2), 143-158.
Edwards, James T. and John A. List (2014): "Toward an Understanding of Why Suggestions Work in Charitable Fundraising: Theory and Evidence from a Natural Field Experiment," *Journal of Public Economics,* 114, 1-13.
Ellingsen, Tore, Magnus Johannesson, Sigve Tjotta, Gaute Torsvik (2010): "Testing Guilt Aversion," *Games and Economic Behavior,* 68, 95-107.
Ellingsen, Tore and Magnus Johannesson (2008): "Anticipated Verbal Feedback Induces Altruistic Behavior," *Evolution and Human Behavior,* 29(2), 100-105.
Engel, Christoph (2011): "Dictator Games: A Meta Study," *Experimental Economics,* 14(4), 583-610.
Engelmann, Dirk and Martin Strobel (2004): "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments," *The American Economic Review,* 94(4), 857-869.
Falk, Armin and Urs Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior,* 54(2), 293-315.
Faravelli, Marco (2007): "How Context Matters: A Survey Based Experiment on Distributive Justice," *Journal of Public Economics,* 91(7-8), 1399-1422.
Fehr, Dietmar (2018): "Is Increasing Inequality Harmful? Experimental Evidence," *Games and Economic Behavior,* 107, 123-134.
Fehr, Ernst, Georg Kirchsteiger and Arno Riedl (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation," *The Quarterly Journal of Economics,* 108(2), 437-459.
Fehr, Ernest and Klaus M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics,* 114(3), 817-868.
Fehr, Ernest and Urs Fischbacher (2004): "Third-Party Punishment and Social Norms," *Evolution and Human Behavior,* 25(2), 63-87.
Feiler, Lauren (2014): "Testing Models of Information Avoidance with Binary Choice Dictator Games," *Journal of Economic Psychology,* 45, 253-267.
Finus, Michael and Pedro Pintassilgo (2013): "The Role of Uncertainty and Learning for the Success of International Climate Agreements," *Journal of Public Economics,* 103, 29-43.
Forsythe, Robert, Joel L. Horowitz, N. E. Savin, and Martin Sefton (1994): "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior,* 6(3), 347-369.
Franzen, Axel, and Sonja Pointner (2013): "The External Validity of Giving in the Dictatorship Game: A Field Experiment Using the Misdirected Letter Technique," *Experimental Economics,* 16(2), 155-159.
Gino, Francesca, Lisa L. Shu, and Max H. Bazerman (2010): "Nameless + Harmless = Blameless: When Seemingly Irrelevant Factors Influence Judgement of (Un)Ethical Behavior," *Organizational Behavior and Human Decision Processes,* 111(2), 93-101.
Gino, Francesca, Michael I. Norton. And Roberto A. Weber (2016): "Motivated Bayesians: Feeling Moral While Acting Egoistically," *Journal of Economic Perspectives,* 30(3), 189-212.
Golman, Russell, David Hagmann, and George Loewenstein (2017): "Information Avoidance," *Journal of Economic Literature,* 55, 96-135.
Green, Stuart P. (2007): "Looting, Law, and Lawlessness," *Tulane Law Review,* 81, 1129-1179.
Grossman, Philip J. and Catherine C. Eckel (2015): "Giving Versus Taking for Cause," *Economic Letters,* 132(C), 28-30.
Grossman, Zachary (2014): "Strategic Ignorance and the Robustness of Social Preferences," *Management Science,* 60(11), 2659-2665.

Grossman, Zachary (2015): "Self-Signaling and Social-Signaling in Giving," *Journal of Economic Behavior and Organization,* 117, 26-39.
Grossman, Zachary and Joël J. van der Weele (2017): "Self-Image and Willful Ignorance in Social Decisions," *Journal of the European Economic Association,* 15, 173-217.
Gurdal, Mehmet Y., Joshua B. Miller, and Aldo Rustichini (2013): "Why Blame?," *Journal of Political Economy,* 121(6), 1205-1247.
Güth, Werner, Steffen Huck, and Wieland Müller (2001): "The Relevance of Equal Splits in Ultimatum Games," *Games and Economic Behavior,* 37, 161-169.
Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982): "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization,* 3(4), 367-388.
Hammann, John R., George Loewenstein, and Roberto A. Weber (2010): "Self-Interest Through Delegation: An Additional Rationale for the Principal-Agent Relationship," *American Economic Review,* 100(4), 1826-1846.
Healy Andrew J., Neil Malhorta, and Cecilia Hyunjung Mo (2010): "Irrelevant Events Affect Voters' Evaluations of Government Performance," *Proceedings of the National Academy of Sciences of the United States of America,* 107(29), 12804-12809.
Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith (1996): "Social Distance and Other-Regarding Behavior in Dictator Games," *The American Economic Review,* 86(3), 653-660.
Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011): "The Online Laboratory: Conducting Experiments in a Real Labor Market," *Experimental Economics,* 14(3), 399-425.
Iriberri, Nagore and Pedro Rey-Biel (2013): "Elicited Beliefs and Social Information in Modified Dictator Games: What do Dictators Believe Other Dictators do?," *Quantitative Economics,* 4(3), 515-547.
Kausel, Edgar E., Santiago Ventura, and Arturo Rodríguez (2019): "Outcome Bias in Subjective Ratings of Performance: Evidence from the (Football) Field," *Journal of Economic Psychology,* 75(B), 1-9.
Kessler, Esther, Maria Ruiz-Martos, and David Skuse (2012): "Destructor Game," *Working Papers,* 11, 1-9.
Khazan, Olga (2020): "Why People Loot," *The Atlantic,* June 2, 2020.
Kimbrough, Erik O. and Alexander Voxtroknutov (2016): "Norms Make Preferences Social," *Journal of the European Economic Association,* 14(3), 608-638.
Konow, James (2000): "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," *American Economic Review*, 90(4), 1072-1091.
Konow, James (2001): "Fair and Square: The Four Sides of Distributive Justice," *Journal of Economic Behavior and Organizations*, 46(2), 137-164.
Konow, James (2005): "Blind Spots: The Effects of Information and Stakes on Fairness Bias and Dispersion," *Social Justice Research,* 18(4), 349-390.
Konow, James (2009): "Is Fairness in the Eye of the Beholder? An Impartial Spectator Analysis of Justice," *Social Choice and Welfare,* 33, 101-127.
Konow, James (2010): "Mixed Feelings: Theories of and Evidence on Giving," *Journal of Public Economics,* 94(3-4), 279-297.
Konow, James (2012): "Adam Smith and the Modern Science of Ethics," *Economics and Philosophy,* 28(3), 333-362.
Konow, James (2019): "Can Ethics Instruction Make Economics Students More Pro-Social?," *Journal of Economic Behavior and Organization,* 166, 724-734.
Konow, James (2022): "Moral Salience and Conditional Altruism: Reconciling Jekyll and Hyde Paradoxes," working paper.
Konow, James, Tatsuyoshi Saijo, and Kenju Akai (2020): "Equity Versus Equality: Spectators, Stakeholders and Groups," *Journal of Economic Psychology,* 77, 1-28.
Korenok, Oleg, Edward L. Millner, and Laura Razzolini (2012): "Are Dictators Averse to Inequality?," *Journal of Economic Behavior and Organization,* 82(2-3), 543-547.
Korenok, Oleg, Edward L. Millner, and Laura Razzolini (2014): "Taking, Giving, and Impure Altruism in Dictator Games," *Experimental Economics,* 17(3), 488-500.

Korenok, Oleg, Edward L. Millner, and Laura Razzolini (2017): "Feelings of Ownership in Dictator Games," *Journal of Economic Psychology,* 61, 145-151.
Korenok, Oleg, Edward L. Millner, and Laura Razzolini (2018): "Taking Aversion," *Journal of Economic Behavior and Organization,* 150, 397-403.
Krawczyk, Michal and Fabrice Le Lec (2010): "'Give Me a Chance!'An Experiment in Social Decision Under Risk," *Experimental Economics,* 13(4), 500-511.
Krupka, Erin L. and Roberto A. Weber (2013): "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?," *Journal of the European Economic Association,* 11(3), 495-524.
Kühl, Leonie and Nora Szech (2017): "Physical Distance and Cooperativeness Towards Strangers," *CESifo Working Paper,* 6825, 1-64.
Larson, Tara and C. Monica Capra (2009): "Exploiting Moral Wiggle Room: Illusory Preference for Fairness? A Comment," *Judgement and Decision Making,* 4(6), 467-474.
Lazear, Edward P., Urlike Malmendier, and Roberto A. Weber (2012): "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics,* 4, 136-163.
Levine, David K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics,* 1, 593-622.
List, John A. (2007): "One the Interpretation of Giving in Dictator Games," *Journal of Political Economy,* 115(3), 482-493.
Mollerstrom, Johanna, Bjorn-Atle Reme and Erik O. Sorensen (2015): "Luck, Choice and Responsibility – An Experimental Study of Fairness Views," *Journal of Public Economics*, 131, 33-40.
Müller, Daniel and Sander Renes (2020): "Fairness Views and Political Preferences: Evidence from a Large and Heterogeneous Sample," *Social Choice and Welfare*, 56, 679-711.
Oexl, Regine and Zachary J. Grossman (2013): "Shifting the Blame to a Powerless Intermediary," *Experimental Economics,* 16(3), 306-312.
Offerman, Theo (2002): "Hurting Hurts More Than Helping Helps," *European Economic Review,* 46(8), 1423-1437.
Oxoby, Robert J. and John Spraggon (2008): "Mine and Yours: Property Rights in Dictator Games," *Journal of Economic Behavior,* 65(3-4), 703-713.
Peterson, Erik, Sharad Goel, and Shanto Iyengar (2019): "Partisan Selective Exposure in Online News Consumption: Evidence from the 2016 Presidential Campaign," *Politcal Science Research and Methods,* 1-17.
Quarantelli, E. L. and Russell R. Dynes (1968): "Looting in Civil Disorders: An Index of Social Change," *The American Behavioral Scientist,* 11, 7-10.
Rabin, Matthew (1993): "Incorporating Fairness into Game Theory and Economics," *The American Economic Review,* 83(5), 1281-1302.
Rey-Biel, Pedro, Roman Sheremeta, and Neslihan Uler (2018): "When Income Depends on Performance and Luck: The Effects of Culture and Information on Giving," *Experimental Economics and Culture (Research in Experimental Economics, vol. 20),* Bingley, UK: Emerald Publishing Ltd, 167-203.
Rigdon, Mary, Keiko Ishii, Motoki Watabe, Shinobu Kitayama (2009): "Minimal Social Cues in the Dictator Game," *Journal of Economic Psychology,* 30(3), 358-367.
Rubin, Jared and Roman Sheremeta (2016): "Principal-Agent Settings with Random Shocks," *Management Science,* 62(4), 985-999.
Serra-Garcia, Marta and Nora Szech (2019): "The (In)Elasticity of Moral Ignorance," *KIT Working Paper Series in Economics,* 134, 1-60.
Sezer, Ovul, Ting Zhang, Francesca Gino, Max H. Bazerman (2016): "Overcoming the Outcome Bias: Making Intentions Matter," *Organizational Behavior and Human Decision Processes,* 137, 13-26.
Shang, Jen and Rachel Croson (2009): "A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods," *The Economic Journal,* 119(540), 1422-1439.

Smith, Alexander (2015): "On the Nature of Pessimism in Taking and Giving Games," *Journal of Behavioral and Experimental Economics,* 54, 50-57.
Sobel, Joel (2005): "Interdependent Preferences and Reciprocity," *Journal of Economic Literature,* 43, 392-436.
Spiekermann, Kai and Arne Weiss (2016): "Objective and Subjective Compliance: A Norm-Based Explanation of 'Moral Wiggle Room'," *Games and Economic Behavior,* 96, 170-183.
Spranca, Mark, Elisa Minsk, and Jonathan Baron (1991): "Omission and Commission in Judgment and Choice," *Journal of Experimental Social Psychology,* 27, 76-105.
Sutter, Matthias, Jürgen Huber, Michael Kirchner, Matthias Stefan, and Markus Walzl (2020): "Where to Look for the Morals in Markets," *Experimental Economics,* 23, 30-52.
Touré-Tillery, Maferima, and Ayelet Fishbach (2017): "Too Far to Help: The Effect of Perceived Distance on the Expected Impact and Likelihood of Charitable Action," *Journal of Personality and Social Psychology,* 112(6), 860-876.
Van Koten, Silvester, Andreas Ortmann, and Vitezslav Babicky (2013): "Fairness in Risky Environments: Theory and Evidence," *Games,* 4(2), 208-242.
Walzer, Michael (1983): "Spheres of Justice: A Defense of Pluralism and Equality," *The Journal of Philosophy,* 83(8), 457-468.
Whitt, Sam and Rick K. Wilson (2007): "The Dictator Game, Fairness and Ethnicity in Postwar Bosnia," *American Journal of Political Science,* 51(3), 655-668.
Williams, Bernard (1981): *Moral Luck*. Cambridge: Cambridge University Press.
Xiao, Erte and Daniel Houser (2009): "Avoiding the Sharp Tongue: Anticipated Written Messages Promote Fair Economic Exchange," *Journal of Economic Psychology,* 30(3) 393-404.
Zhang, Le, and Andreas Ortmann (2013): "On the Interpretation of Giving, Taking, and Destruction in Dictator Games and Joy-of-Destruction Games," Australian School of Business Research Paper No. 2012ECON50A (http://dx.doi.org/10.2139/ssrn.219040).
Zizzo, Daniel John and Andrew J. Oswald (2001): "Are People Willing to Pay to Reduce Others' Income," *Annales d'Économie et de Statistique,* 63, 39-65.
Zizzo, Daniel John (2003): "Money Burning and Rank Egalitarianism with Random Dictators," *Economic Letters,* 81(2), 263-266.
Zizzo, Daniel John (2010): "Experimenter Demand Effects in Economic Experiments," *Experimental Economics,* 13, 75-98.

Appendix 1: Proofs

Note below that primes denote derivatives and, in the case of g, partial derivatives with respect to x, e.g., $g' \equiv {}^{\partial g}/_{\partial x}$ and $g'' \equiv {}^{\partial^2 g}/_{\partial x^2}$

Proof of Theorem 3.1

$${}^{dU}/_{dz} = -u'(x+Z-z) + \sigma\phi f'\left(z - \tfrac{Z}{2} + \tfrac{X}{2} - x - \theta r(x-\tilde{x})\tfrac{X}{2}\right) + \sigma g'(x,\alpha) = 0.$$

Substituting $z(x)$ and differentiating,

$$-u'' + u''\frac{dz}{dx} + \sigma\phi f''\frac{dz}{dx} - \sigma\phi f'' - \sigma\phi\theta f'' r'\frac{X}{2} + \sigma g''\frac{dz}{dx} = 0,$$

$${}^{dz}/_{dx} = \frac{u''+\sigma\phi f''+\sigma\phi\theta f'' r' {}^{X}/_{2}}{u''+\sigma\phi f''+\sigma g''}.$$

If x is randomly assigned, then $r = 0$ and $0 < \frac{dz}{dx} < 1$.

Otherwise, $r' > 0$ and sanctioning increases the value of ${}^{dz}/_{dx}$.

Proof of Theorem 3.2

$${}^{dU}/_{dz} = \sigma\phi f'\left(z - \frac{Z}{2} + \frac{X}{2} - x - \theta r(x-\tilde{x})\cdot\frac{X}{2}\right) = 0$$

$$\Longrightarrow\ z = \frac{Z}{2} - \frac{X}{2} + x + \theta r(x-\tilde{x})\frac{X}{2}$$

$${}^{dz}/_{dx} = 1 + \theta r'(x-\tilde{x})\tfrac{X}{2}$$

which equals 1, if x is randomly assigned and $r = 0$, and is greater than 1, if x is chosen and $r' > 0$. Note also that

${}^{d^2 z}/_{dx^2} = \theta r''\frac{X}{2} < 0 \implies$ asymmetric sanctioning.

Proof of Theorem 4.1

Define salience in the reference state to be at the level of a given first stage sinner and saints game, $\tilde{\sigma}$. Moreover, suppose, as usual, that $x^L > \underline{\gamma}$ and $x^H < \bar{\gamma}$. Then

$$\hat{\gamma}(x) = \begin{cases} \int_{\gamma^H}^{\bar{\gamma}} \gamma\rho(\gamma)d\gamma \Big/ \int_{\gamma^H}^{\bar{\gamma}} \rho(\gamma)d\gamma \ \ if\ x = x^H \\ \qquad x(\tilde{\sigma})\ if\ x^L < x < x^H \\ \int_{\underline{\gamma}}^{\gamma_L} \gamma\rho(y)d\gamma \Big/ \int_{\underline{\gamma}}^{\gamma_L} \rho(y)d\gamma \ \ if\ x = x^L \end{cases}$$

The first order condition is

$${}^{dU}/_{dz} = \sigma\phi f'\left(z - \eta_z - \theta r(\hat{\gamma}-\tilde{\gamma})\cdot\frac{Z}{2}\right) = 0$$

$$\Rightarrow z = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r(\hat{\gamma}-\tilde{\gamma})\cdot\frac{Z}{2}$$

When $x^L < x < x^H$, $\hat{\gamma} = x$, and

${}^{dz}/_{dx} = 1 + \theta r'(\hat{x}-\tilde{x})\ \frac{Z}{2}$, and

$$d^2z/_{dx^2} = \theta r'' \frac{Z}{2} < 0.$$

When $x = x^L$,

$$\tilde{z}\Big|_{x^L} = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r(\hat{\gamma}(x^L) - \tilde{\gamma})\frac{Z}{2}$$
$$< \tilde{z}\Big|_{\gamma^L} = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r\,(\gamma^L - \tilde{\gamma})\,\frac{Z}{2}$$

since $\tilde{\gamma}(x^L) < \gamma^L$.

When $x = x^H$,

$$\tilde{z}\Big|_{x^H} = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r(\hat{\gamma}(x^H) - \tilde{\gamma})\frac{Z}{2}$$
$$> \tilde{z}\Big|_{\gamma^H} = \frac{Z}{2} - \frac{X}{2} - \frac{Y}{2} + x + \theta r(\gamma^H - \tilde{\gamma})\frac{Z}{2}$$

since $\hat{\gamma}(x^H) > \gamma^H$. This implies $\tilde{z}$ is concave in x with a discontinuous decrease at x^L and a discontinuous increase at x^H. From Konow (2022), we have Assumption 7, $dn/_{dx^L} < 0$, Definition 1, $\partial\sigma/_{\partial n} < 0$, and by Theorem 2.2, $dx/_{d\sigma} > 0$. Then if a D gives $\tilde{\gamma}$ at the reference salience $\tilde{\sigma}$, then increasing x^L increases the threshold transfer in the sinners and saints game:

$$\partial\tilde{x}/_{\partial x^L} = \frac{dx}{d\sigma}\frac{\partial\sigma}{\partial n}\frac{dn}{x^L} > 0$$

for interior solutions. Note that $x(\tilde{\gamma}, \sigma)$, so if x is held constant,

$$dx = \frac{dx}{d\hat{\gamma}}d\hat{\gamma} + \frac{\partial x}{\partial\sigma}d\sigma = 0$$
$$\Rightarrow \frac{d\hat{\gamma}}{d\sigma}\Big|_x = \frac{\partial x/_{\partial\sigma}}{\partial x/_{\partial\hat{\gamma}}} = -\frac{\partial x}{\partial\sigma} < 0 \text{ since } \frac{\partial x}{\partial\hat{\gamma}} = 1$$

for interior solutions. Then

$$\frac{\partial\hat{\gamma}}{\partial x^L}\Big|_x = \frac{d\hat{\gamma}}{d\sigma}\frac{\partial\sigma}{\partial n}\frac{dn}{dx^L} < 0$$

and $\frac{\partial z}{\partial x^L} = \frac{\partial z}{\partial\hat{\gamma}}\frac{\partial\hat{\gamma}}{\partial x^L} < 0$ since $\frac{\partial z}{\partial\hat{\gamma}} > 0$.

Proof of Theorem 4.2

$$U = u(\bar{z}) + \sigma\phi f\left(z - \frac{Z}{2} + \eta_x(x^L) - x - \theta r\big(x - \tilde{x}(x^L)\big)\cdot\frac{Z}{2}\right)$$
$$dU/_{dz} = \sigma\phi f'\left(z - \frac{Z}{2} + \eta_x(x^L) - x - \theta r\big(x - \tilde{x}(x^L)\big)\cdot\frac{Z}{2}\right) = 0$$
$$\Rightarrow z = \frac{Z}{2} - \eta_x(x^L) + x + \theta r\big(x - \tilde{x}(x^L)\big)\cdot\frac{Z}{2}))$$
$$\frac{dz}{dx^L}\Big|_x = -\frac{d\eta_x}{dx^L} - \theta r'\big(x - \tilde{x}(x^L)\big)\cdot\frac{Z}{2}\cdot\frac{d\tilde{x}}{dx^L} < 0.$$

Proof of Theorem 5.1

Since $\beta = 0$,

$$U = u(x_r) + \sigma\phi f\left(z - \tfrac{1+b}{2}Z - \theta r(\hat{\gamma} - \tilde{\gamma})\cdot x'\right).$$

$$dU/_{dz} = \sigma\phi f'\left(z - \frac{1+b}{2}Z - \theta \cdot r(\hat{\gamma} - \tilde{\gamma}) \cdot x'\right) = 0$$
$$\Longrightarrow z = \tfrac{1+b}{z}Z + \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot x'.$$

Distributive preferences imply the fixed amount $\frac{1+b}{z}Z$. The fact that $\hat{\gamma}^f > \hat{\gamma}^u$ implies the corresponding $z^f > z^u$ for a given value of x'.

Proof of Theorem 5.2

Let z^{x,x_r} be the R's choice of $\tilde{z}$ for the D's choice x and the realization x_r.

$z^{fF} = \frac{1+b}{z}Z + \theta r(\hat{\gamma}^f - \tilde{\gamma}) \cdot F$, since $x' = F$

$z^{uU} = \frac{1+b}{z}Z + \theta r(\hat{\gamma}^u - \tilde{\gamma}) \cdot F$, since $x' = F$

$z^{fU} = \frac{1+b}{z}Z + \theta r(\hat{\gamma}^f - \tilde{\gamma}) \cdot (qF + (1-q)L)$, since $x' = EX^f = qF + (1-q)L$

$z^{uF} = \frac{1+b}{z}Z + \theta r(\hat{\gamma}^u - \tilde{\gamma}) \cdot \big((1-q)F + qL\big)$ since $x' = EX^u = (1-q)F + qL$

$z^{fF} > z^{fU}$ as long as $z^{fU} < Z$ since $r > 0$ and $F > qF + (1-q)L$

$z^{uU} < z^{uF}$ since $r < 0$ and $F > (1-q)F + qL$

If D chooses with certainty, then for Rs

$$U = \sigma f(\phi\left(z - \frac{1+b}{2}Z - \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot F\right)$$
$$\Longrightarrow z = \tfrac{1+b}{2}Z + \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot F.$$

If D chooses fair,

$z^F = \frac{1+b}{z}Z + \theta r(\hat{\gamma}^f - \tilde{\gamma}) \cdot F = z^{fF}$.

If D chooses unfair,

$z^U = \frac{1+b}{z}Z + \theta r(\hat{\gamma}^u - \tilde{\gamma}) \cdot F = z^{uU}$.

Proof of Theorem 6.1

The proof proceeds by reducing the set of optimal choices by showing that some are dominated, and identifies thresholds for certain choices using the assumption that fairness is homogeneous in ϕ. A partial ranking of more or less generous choices can then be identified.

6.1.1. *NRA* dominates *NRB*

$$EU(NRA) = u(H) + .5\sigma^l\phi f(L-F) + .5\sigma^l g(L,\alpha) + .5\sigma^l g(F,\alpha)$$
$$> EU(NRB) = u(F) + .5\sigma^l\phi f(L-F) + .5\sigma^l g(L,\alpha) + .5\sigma^l g(F,\alpha)$$

since $u(H) > u(F)$.

6.1.2. *R2A* dominates *R2B*

$$U(R2A) = u(H) + \sigma^m g(F,\alpha) > U(R2B) = u(F) + \sigma^m\phi f(L-F) + \sigma^m g(L,\alpha)$$

for all altruistic and selfish Ds. This inequality is reversed only if we relax Assumption 2 about altruism being second order and D is so spiteful (denoted $\alpha \ll 0$) that $g(F,\alpha) < \frac{1}{\sigma^m}[u(F) - u(H)] + g(L,\alpha) < 0$.

6.1.3. Fairest Ds choose Reveal and B in state 1 (R1B)

Note R2A dominates R2B if reveal and state 2 obtains, and NRA dominates NRB, if do not reveal. So we compare the expected utility of reveal and B in state 1 and A in state 2 to NRA:

$$EU(R1B, R2A) = .5U(R1B) + .5U(R2A) > EU(NRA)$$
$$.5u(F) + .5\sigma^m g(F,\alpha) + .5u(H) + .5\sigma^m g(F,\alpha)$$
$$> u(H) + .5\sigma^l \phi f(L-F) + .5\sigma^l g(L,\alpha) + .5\sigma^l g(F,\alpha)$$
$$\Rightarrow \phi > \phi^{R1B} \equiv \frac{\frac{1}{\sigma^l}[u(H) - u(F)] + \left(\frac{\sigma^l - 2\sigma^m}{\sigma^l}\right) g(F,\alpha) + g(L,\alpha)}{-f(L-F)} > 0$$

under Assumption 2.

6.1.4. Least fair Ds choose Reveal and A in state 1 (R1A)

Compare as above, but assume D chooses A if reveals and state 1 obtains.

$$EU(R1A, R2A) = .5U(R1A) + .5U(R2A) > EU(NRA)$$
$$u(H) + .5\sigma^m \phi f(L-F) + .5\sigma^m g(L,\alpha) + .5\sigma^m g(F,\alpha)$$
$$> u(H) + .5\sigma^l \phi f(L-F) + .5\sigma^l g(L,\alpha) + .5\sigma^l g(F,\alpha)$$
$$\Rightarrow \phi < \phi^{R1A} \equiv \frac{g(F,\alpha) + g(L,\alpha)}{-f(L-F)} > 0$$

for the mean D and on average from assumptions about the distribution of altruism types.

6.1.5. Fairer Ds choose 1B over 1A

If, in the baseline for state 1, Ds prefer B:

$$U(1B) > U(1A)$$
$$u(F) + \sigma^h g(F,\alpha) > u(H) + \sigma^h \phi f(L-F) + \sigma^h g(L,\alpha)$$
$$\Rightarrow \phi > \phi^{1B} \equiv \frac{\frac{1}{\sigma^h}[u(H) - u(F)] - g(F,\alpha) + g(L,\alpha)}{-f(L-F)} > 0$$

by Assumption 2.

6.1.6. $2A$ dominates $2B$

The proof for the baseline for state 2 parallels that for 6.1.2 but substituting σ^h for σ^m:

$$\phi > \phi^{2A} \equiv \frac{\frac{1}{\sigma^h}[-u(H) + u(F)] - g(F,\alpha) + g(L,\alpha)}{-f(L-F)} \leq 0$$

such that all Ds choose 2A unless we relax Assumption 2 and D is so spiteful ($\alpha \ll 0$) that $g(F,\alpha) < \frac{1}{\sigma^h}[-u(H) + u(F)] + g(L,\alpha)$<0.

Rankings by generosity

Generosity in the information game can be estimated in ordinal terms by inferring optimal choices according to intrinsic moral motivation. Specifically, the primary ranking is based on values of ϕ for two reasons: the focus on fairness in norm avoidance and the assumed second order effects of altruism. Moreover, recall the conclusions that transfers are increasing in α and ϕ and the assumption that $Cov(\alpha, \phi) > 0$, which imply that a higher ϕ implies greater generosity, ceteris paribus. An exception is super-fair choices where generosity is decreasing in ϕ, but such choices are ruled out by design in this game where the recipient can never receive

more than F. Specifically, the following rankings of expected generosity follow from the threshold values of ϕ that are derived above:

$\hat{\gamma}^{R1B} > \hat{\gamma}^{NRA} > \hat{\gamma}^{R1A}$ from 6.1.1, 6.1.3 and 6.1.4.

$\hat{\gamma}^{R1B} > \hat{\gamma}^{R2A} > \hat{\gamma}^{R1A}$ from the fact that those choosing R2A if state 1 obtains are a mixture of R1A and R1B types.

$\hat{\gamma}^{1B} > \hat{\gamma}^{1A}$ from 6.1.5.

Note that R2B and 2B are predicted to be dominated under Assumption 2. For the purposes later of analyzing sanctioning behavior of these choices, we identify what preferences could result in these choices, even if the negligible incidence of such choices justifies Assumption 2 and its implications here. Specifically, R2B and 2B would only be chosen, if some Ds are so spiteful that their spite swamps their material utility, which leads to the following rankings:

$\hat{\gamma}^{R2A} > \hat{\gamma}^{R2B}$

$\hat{\gamma}^{2A} > \hat{\gamma}^{2B}$

Proof of Corollary 6.1

If the entitlement is an equal split of final payoffs, fairer Ds prefer NRB over NRA if

$$EU(NRB) > EU(NRA)$$

$$u(F) + .5\sigma^l g(F,\alpha) + .5\sigma^l \phi f\left(L - \tfrac{F+L}{2}\right) + .5\sigma^l g(L,\alpha)$$

$$> u(H) + .5\sigma^l \phi f\left(L - \tfrac{H+L}{2}\right) + .5\sigma^l g(L,\alpha) + .5\sigma^l \phi f\left(L - \tfrac{H+F}{2}\right) + .5\sigma^l g(F,\alpha)$$

$$\Rightarrow \phi > \phi^{NRB} \equiv \frac{\frac{2}{\sigma^l}[u(H) - u(F)] - g(F,\alpha)}{f\left(\frac{L-F}{2}\right) - f\left(\frac{L-H}{2}\right) - f\left(\frac{F-H}{2}\right)} > 0$$

by Assumption 2 and since $\left|f\left(\tfrac{L-F}{2}\right)\right| < -f\left(\tfrac{L-H}{2}\right) - f\left(\tfrac{F-H}{2}\right)$.

$\Rightarrow \hat{\gamma}^{NRB} > \hat{\gamma}^{NRA}$

Proof of Theorem 6.2

The utility function of a spectator is

$$U = u(\bar{z}) + \sigma\phi f(z - \eta_z - \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot x')$$

where $\eta_z = 0$ since $\beta = 0$ and $z \leq 0$.

$$\frac{dU}{dZ} = \sigma\phi f'(z - \theta r(\hat{\gamma} - \tilde{\gamma})) \cdot x') = 0$$

$$\Rightarrow \tilde{z} = \theta r(\hat{\gamma} - \tilde{\gamma}) \cdot x'$$

When the payoffs of choices are known, the scale is entitlement, which is assumed to be *F*, i.e., $x' = F$. Ds who choose not to reveal are making an unfair choice in expectations, since they could always reveal and guarantee that R receives F. In Not reveal, the scale depends on the combination of choice and outcome as stated in section 5 on outcome bias. Specifically, if the unfair outcome for R obtains (i.e., L), then choice matches outcome, and the scale is R's entitlement, F, which is the case for option A if state 1 obtains (NRA1) and option B in state 2 (NRB2). If the fair outcome obtains, then choice does not match outcome, and the scale is the expected payoff to R, viz., $x' = EX^u = .5F + .5L \equiv EX < F$, which is the case for option A in

state 2 (NRA2) and option B in state 1 (NRB1). Note that choices R1B and 1B are the only ones that produce equal payoffs F and should not be punished, since the threshold for generosity, $\tilde{\gamma}$, cannot exceed F (indeed, they would be rewarded by some spectators for whom $\tilde{\gamma} < F$ were it not for the constraint by design that sanctions are non-positive). Then some spectators will punish choices other than R1B and 1B with the ideal sanction varying according to γ and x' as follows:

$$\tilde{z}^{R1A} = \theta r(\hat{\gamma}^{R1A} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{R1B} = \theta r(\gamma^{R1B} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{R2A} = \theta r(\hat{\gamma}^{R2A} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{R2B} = \theta r(\gamma^{R2B} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{NRA1} = \theta r(\hat{\gamma}^{NRA1} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{NRA2} = \theta r(\gamma^{NRA2} - \tilde{\gamma}) \cdot EX$$
$$\tilde{z}^{NRB1} = \theta r(\hat{\gamma}^{NRB1} - \tilde{\gamma}) \cdot EX$$
$$\tilde{z}^{NRB2} = \theta r(\gamma^{NRB2} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{1A} = \theta r(\hat{\gamma}^{1A} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{1B} = \theta r(\gamma^{1B} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{2A} = \theta r(\hat{\gamma}^{2A} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}^{2B} = \theta r(\gamma^{2B} - \tilde{\gamma}) \cdot F$$

From inspection of these equations and the rankings of estimated generosity of choices from Theorem 6.1 and Corollary 6.1, we have the following rankings of the ideal sanction, $\tilde{z}^{C\omega}$, according to choice, *C*, and, in the case of NRA and NRB, realized state, $\omega \in \{1,2\}$:

$$0 = \tilde{z}^{R1B} > \tilde{z}^{NRA2} > \tilde{z}^{NRA1} > \tilde{z}^{R1A}$$
$$0 = \tilde{z}^{R1B} > \tilde{z}^{R2A} > \tilde{z}^{R1A}$$
$$\tilde{z}^{R2A} > \tilde{z}^{R2B}$$
$$\tilde{z}^{NRB1} > \tilde{z}^{NRA2} > \tilde{z}^{NRA1}$$
$$\tilde{z}^{NRB1} > \tilde{z}^{NRB2} > \tilde{z}^{NRA1}$$
$$0 = \tilde{z}^{1B} > \tilde{z}^{1A}$$
$$\tilde{z}^{2A} > \tilde{z}^{2B}$$

Proof of Theorem 7.1

In the baseline, salience is high, σ^h, and the D chooses fair if

$U(F) > U(U)$

$$u(F) + \sigma^h 3g(F,\alpha) > u(H) + \sigma^h \phi f(H-F) + \sigma^h \phi 2f(L-F) + \sigma^h g(H,\alpha) + \sigma^h 2g(L,\alpha)$$
$$\phi > \phi^{BF} \equiv \frac{\frac{1}{\sigma^h}[u(H) - u(F)] + [g(H,\alpha) + 2g(L,\alpha)] - 3g(F,\alpha)}{-f(H-F) - 2f(L-F)}$$

In the delegation game, the D chooses to allocate fairly with salience σ^m over delegating with salience σ^l if

$U(F) > EU(D)$

$$u(F) + \sigma^m 3g(F,\alpha) > qu(F) + (1-q)u(H) + (1-q)\sigma^l \phi f(H-F) + (1-q)\sigma^l \phi 2f(L-F) + q\sigma^l 3g(F,\alpha) + (1-q)\sigma^l g(H,\alpha) + (1-q)\sigma^l 2g(L,\alpha)$$

$$\phi > \phi^{AF} \equiv \frac{\frac{1}{\sigma^l}[u(H) - u(F)] + [g(H,\alpha) + 2g(L,\alpha)] - \frac{\sigma^m - q\sigma^l}{\sigma^l(1-q)} \cdot 3g(F,\alpha)}{-f(H-F) - 2f(L-F)}$$

In the delegation game, the D chooses to allocate unfairly over delegating if

$\underline{U(U) > EU(D)}$

$$u(H) + \sigma^m \phi f(H-F) + \sigma^m \phi 2f(L-F) + \sigma^m g(H,\alpha) + \sigma^m 2g(L,\alpha)$$
$$> qu(F) + (1-q)u(H) + (1-q)\sigma^l \phi f(H-F) + (1-q)\sigma^l \phi 2f(L-F)$$
$$+ q\sigma^l 3g(F,\alpha) + (1-q)\sigma^l g(H,\alpha) + (1-q)\sigma^l 2g(L,\alpha)$$

$$\phi < \phi^{AU} \equiv$$
$$\frac{\frac{1}{\sigma^l + \frac{1}{q}(\sigma^m - \sigma^l)}[u(H) - u(F)] + [g(H,\alpha) + 2g(L,\alpha)] - \frac{q\sigma^l}{\sigma^m - (1-q)\sigma^l} \cdot 3g(F,\alpha)}{-f(H-F) - 2f(L-F)}$$

Under Assumption 2, we disregard altruism and focus on the other terms. Note $\phi^{AF} > \phi^{AU}$ since $\sigma^l < \sigma^l + \frac{1}{q}(\sigma^m - \sigma^l)$, so Ds with $\phi^{AF} > \phi > \phi^{AU}$ delegate. Note $\phi^{AF} > \phi^{BF}$ since $\sigma^l < \sigma^h$, so fewer Ds allocated fairly in the delegation game than in the dictator game.

The thresholds above establish the following rankings of the estimated generosity of dictators across cases:

$\hat{\gamma}_D^{AF} > \hat{\gamma}_D^{BF} > \hat{\gamma}_D^{D} > \hat{\gamma}_D^{AU}$,
$\hat{\gamma}_D^{D} > \hat{\gamma}_D^{BU}$,

where $\hat{\gamma}_D^{AF} > \hat{\gamma}_D^{BF}$ and $\hat{\gamma}_D^{D} > \hat{\gamma}_D^{BU}$ by $\phi^{AF} > \phi^{BF}$ and $\hat{\gamma}_D^{BU} > \hat{\gamma}_D^{AU}$ by $\phi^{AF} > \phi^{AU}$.

Finally, the intermediary chooses fair if

$\underline{U(F) > U(U)}$

$$u(F) + \sigma^I 3g(F,\alpha)$$
$$> u(H) + \sigma^I \phi f(H-F) + \sigma^I \phi 2f(L-F) + \sigma^I g(H,\alpha) + \sigma^I 2g(L,\alpha)$$

$$\phi > \phi^I \equiv \frac{\frac{1}{\sigma^I}[u(H) - u(F)] + [g(H,\alpha) + 2g(L,\alpha)] - 3g(F,\alpha)}{-f(H-F) - 2f(L-F)}$$

For Is, the threshold value for choosing fair, ϕ^I, is solved the same as for Ds in the baseline (ϕ^{BF}) except for the replacement of σ^h with $\sigma^I \in [\sigma^l, \sigma^m]$. Disregarding altruism terms, $\sigma^I = \sigma^l \Rightarrow \phi^I = \phi^{AF}$, and $\sigma^I = \sigma^m \Rightarrow \phi^I > \phi^{AU}$ since

$\sigma^m > \sigma^l + \frac{1}{q}(\sigma^m - \sigma^l)$.

These imply the following ranking of the estimated generosity of intermediaries:

$\hat{\gamma}_D^{AF} \geq \hat{\gamma}_I^{DF} > \hat{\gamma}_I^{DU} > \hat{\gamma}_D^{AU}$.

$\underline{\text{Proof of Theorem 7.2}}$

From the proof to Theorem 6.2, the spectator's ideal sanction is

$\tilde{z}_P^{GC} = \theta r(\hat{\gamma}_P^{GC} - \tilde{\gamma}) \cdot x'$

applying the notation for $\hat{\gamma}$ in the delegation game to this equation.

When fair is chosen directly or the subject makes no choice, the theory of virtue preferences predicts no punishment, namely, in the following cases:

$\tilde{z}_D^{BF} = \tilde{z}_I^{BF} = \tilde{z}_D^{AF} = \tilde{z}_I^{AF} = \tilde{z}_I^{DF} = \tilde{z}_I^{BU} = \tilde{z}_I^{AU} = 0$

For the remaining five cases, the ideal punishments are:

$$\tilde{z}_D^{DF} = \theta r(\hat{\gamma}_D^D - \tilde{\gamma}) \cdot EX^u$$
$$\tilde{z}_D^{BU} = \theta r(\hat{\gamma}_D^{BU} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}_D^{AU} = \theta r(\hat{\gamma}_D^{AU} - \tilde{\gamma}) \cdot F$$
$$\tilde{z}_D^{DU} = \theta r(\hat{\gamma}_D^D - \tilde{\gamma}) \cdot F$$
$$\tilde{z}_I^{DU} = \theta r(\hat{\gamma}_I^{DU} - \tilde{\gamma}) \cdot F.$$

Inspecting these equations, comparing estimated generosity and noting $EX^u < F$, sanctions can be ranked as follows:

$$\tilde{z}_D^{DF} > \tilde{z}_D^{DU} > \tilde{z}_D^{AU}, \tilde{z}_D^{DU} > \tilde{z}_D^{BU}, \tilde{z}_I^{DU} > \tilde{z}_D^{AU}.$$

Note that delegating is fairer than directly allocating unfairly, making it the most generous, $\hat{\gamma}_D^D$, of the unfair choices. The location of the generosity threshold is an empirical question, but if $\tilde{\gamma} \leq \hat{\gamma}_D^D$, then $\tilde{z}_D^{DF}$ and $\tilde{z}_D^{DU}$ will not be punished.

Appendix 2: Not for Publication
Composite Protocol for Sinners and Saints Experiment

Key: The text below is common to all subjects except where indicated as being specific to subject <X>, [Y] or {Z}. Notes or comments about the protocol that were not in the instructions are in *(parentheses and italicized)*.

Math Test

Instructions

Participants can earn a bonus in this experimental study for making decisions about the allocation of real money. It is important, therefore, that all decision-makers first pass a test consisting of three short questions involving addition and subtraction. You have at most two attempts at each question to get the correct answer and to proceed to the experiment.

Please read the question below, fill in the blank with the correct answer, and then press "Proceed" to go to the next question.

Question 1

Please complete the following subtraction problem:

$$\begin{array}{r} 43 \\ -\ 19 \\ \hline \square \end{array}$$

Question 2

The sum of Columns 1 and 2 on Line 1 below equals −5, as seen in the final Column for Line 1. The sum of Columns 1 and 2 on Line 2 below equals +16, as seen in the final Column for Line 2. Please fill in the blanks under Column 2 with numbers such that these two equations are satisfied.

	Column 1		Column 2		Sum of Cols. 1 + 2
Line 1	−7	+	☐	=	−5
Line 2	+5	+	☐	=	+16

Question 3

Below are two equations involving two numbers, which we will call A and B. Equation 1 involves the sum of these two numbers and Equation 2 the difference between the two numbers. Please fill in the blanks below with the numbers, A and B, such that both equations are true.

	Equation 1	Equation 2
	A	A
	+ B	− B
Total	62	10

The value of A is ☐

The value of B is ☐

General Instructions

Task
This is an experimental study of decision-making involving the allocation of real money. In addition to a reward of $2, which everyone receives, participants can earn a bonus, the amount of which can depend on decisions in the experiment.

Random Assignment
Participants will be randomly assigned to different types and will be randomly matched with other participants.

Anonymity
Participants will remain anonymous, that is, decisions and payments will be private, and no participant will ever be told the identity of any other person with whom he or she is matched.

Now we will go over the instructions that explain your participant type and your decision.

Instructions for Person <X> [Y] {Z}

Quiz about Instructions
Please make sure you read these instructions carefully, because afterwards you must answer some questions about them {in order to continue with the experiment. You will have at most two attempts to answer all questions correctly, or your participation will be terminated}.

Groups
You have been randomly assigned to a group consisting of participants called Persons <X> [Y] {Z}. Other participants have been randomly assigned to groups consisting of participants called Persons <Y> [X] {X and Persons Y}. Each Person [<X>] {Z} is randomly matched with [<one Person Y>] {a Pair consisting of one Person X and one Person Y}.

{Z Payment
For making a decision about X and Y, each Person Z will receive a fixed bonus of $5 *($2 in the Benevolent dictator treatment)*, which is in addition to the reward of $2. The amount of Person Z's bonus has nothing to do with the eventual payments received by Persons X and Y.}

X and Y Endowments
Each X/Y Pair has been provisionally allocated a certain number of points we will call "endowments," which are in addition to their reward of $2 each. Specifically, Person X is endowed with 15 points and Person Y is endowed with 5 points, whereby each point is always worth $0.20 in this experiment. The difference in X and Y endowments is completely arbitrary: in other words, participants are randomly assigned to be either Person X or Person Y, and the difference in their endowments has nothing to do with any other differences between them.

(Double Dictator Treatments
continues below with Instructions for Person <X> [Y] {Z})

Decision of Person X

<As Person X, your decision> {[The decision of Person X]} is to choose how many points to transfer between Persons X and Y. {Your decision, as Person Z, will be described momentarily.} Person Y makes no decisions. Specifically, (you) {[Person X]} may:

(the following appears only where negative transfers are possible, where L is the most negative number)

- Transfer 1 to L points **from Y to X**.
- That is, <you> {[Person X]} may choose a **negative** transfer to Y of –1 to –L points.

(in some cases, only 1 negative point can be transferred rather than a range of points, in which case above it reads "Transfer 1 point from Y to X")

OR

(the following appears where zero transfers are possible in combination with positive and/or negative transfers.)

- Leave the points of X and Y **unchanged**.
- That is, <you> {[Person X]} may choose a transfer to Y of **0** points.

OR

(the following appears only where positive transfers are possible, where H is the highest positive number)

- Transfer 1 to H points **from X to Y**.
- That is, <you> {[Person X]} may choose a **positive** transfer to Y of +1 to +H points.

When you have understood these instructions, click "Continue" below to proceed to the quiz about these instructions.

Quiz about Instructions

We now need to make sure that you have understood the instructions thus far. Please carefully select one answer to each of the three questions below. You have at most two chances to answer all questions correctly, or your participation will be terminated. When you are satisfied with your answers, click "Submit" below.

Question 1

Which of the following statements about endowments is true?

A. Participants are assigned to be either Person X or Person Y based on their relative performance on a task.
B. Person X is endowed with 5 points and Person Y with 15 points.
C. Persons X and Y are randomly assigned to their types, and the difference in their endowments has nothing to do with any other differences between them.

Question 2

What was the value of the **lowest** (or minimum) possible transfer **from X to Y**?

Choose *(This runs from -5 to +5)*

Question 3

What was the value of the **highest** (or maximum) possible transfer **from X to Y**?

Choose *(This runs from -1 to +15)*

<Your> [Person X's] [<Decision>]

<Here again are the transfers you may choose for your decision. You may:

- Transfer 1 to L points **from Y to X**.
- That is, you may choose a **negative** transfer to Y of –1 to –L points.

(in some cases, only 1 negative point can be transferred rather than a range of points, in which case above it reads "Transfer 1 point from Y to X")

OR

(the following appears where zero transfers are possible in combination with positive and/or negative transfers.)

- Leave the points of X and Y **unchanged**.
- That is, you may choose a transfer to Y of **0** points.

OR

(the following appears only where positive transfers are possible, where H is the highest positive number)

- Transfer 1 to H points **from X to Y**.
- That is, you may choose a **positive** transfer to Y of +1 to +H points.

Click on a number in the pull-down menu below in the field marked "Choose" to try out different values for your transfer. Below this, the Points after transfer will update for Persons X and Y. You may revise your choice at any time before submitting it. Once you are satisfied with your choice, click "Submit" below.>

[Below you can see a summary of the Endowments, the transfer chosen by Person X, and the Points after transfer. Please click "Continue" below to see information about an additional allocation.]

	[<Person X	Person Y
Endowments	15	5>]
<I transfer> [Person X transferred] [<this amount		Choose
Points after transfer>]		

{Further Instructions for Person Z

Quiz about Further Instructions

Please make sure you read these instructions carefully, because afterwards you must answer some questions about them in order to continue with the experiment. You will have at most two attempts to answer all questions correctly, or your participation will be terminated.

Decision of Person Z

Your decision as Person Z is to allocate 40 additional points between Persons X and Y. These 40 points will be added to the amounts X and Y receive after the transfer Person X makes from the initial endowments, which total 20 points (15 to X and 5 to Y). You will be told an amount that

Person X chooses to transfer to Person Y, and then you will allocate an amount out of the 40 additional points to Person X. Any remaining points out of the 40 that you do not allocate to Person X will go to Person Y, so no points are lost.

Note that, at the time Person X makes their decision, Person X does not yet know about the 40 additional points or about the existence of Person Z or of any decision by Z. At the end of the experiment after all decisions have been made, however, the other participants will be told about Person Z and the decision by Z that applies to their payments.

Example of Z Choice
We will now go through an Example in order to familiarize you with your decision. We do not yet know which transfer Person X will choose, but, for this Example, suppose that X chooses to transfer +1 (or -1 if +1 is not an option in that treatment) point to Y. That is, suppose X chooses to transfer 1 point from X to Y (or transfer 1 point from Y to X). Your decision is to choose how to allocate the additional 40 points given this X transfer.

The table below summarizes for this Example a Transfer chosen by X and the Points of X and Y after X's transfer. You choose how many, if any, of the 40 additional points you wish to allocate to X. You may try out different values from 0 to 40 to allocate to X from the pull-down menu in the field below marked "Choose." The amount you choose will be added to Person X's points, and the remaining amount from the 40 additional points will be added to Person Y's points, so no points are lost. On the far right, you will see the Total Points to X and Y after X's transfer and after your allocation of the 40 points. This is only an Example, and you can change your answer later.

Suppose Person X transfers +1 (or -1) points to Y	**Points after X's transfer**		**I allocate this amount to Person X**	**Total points**	
	X	**Y**		**X**	**Y**

Multiple Choices by Person Z
Since Person X's transfer is not yet available, you need to make multiple choices: you decide how many of the additional 40 points to allocate to X for each possible transfer Person X can make to Person Y. That is, your decision is to choose allocations to X from the 40 points assuming X makes transfers to Y of –L points, -L+1 points, -L+2 points, etc. Later, when X's choice about how much to transfer to Person Y is available, it will be matched with your corresponding choice about the allocation of the additional 40 points. The final bonus payments to X and Y are based on their Endowments adjusted for X's transfer plus the points you allocate in your corresponding choice.

When you have understood these instructions, click "Continue" below to proceed to the quiz about these instructions.

Quiz about Further Instructions

We now need to make sure that you have understood these further instructions. Please carefully select one answer to each of the three questions below. You have at most two chances to answer all questions correctly, or your participation will be terminated. When you are satisfied with your answers, click "Submit" below.

Question 1
Which of the following statements is true?

A. Any of the 40 points Person Z does not allocate to Person X are lost.
B. The payments of Persons X and Y will be based on a randomly chosen allocation of Person Z.
C. Person Z makes multiple choices: Z chooses how many of 40 additional points to allocate to Person X for each possible transfer Person X can make to Y.

Question 2
Which of the following statements is true?

A. The maximum number of points any Person X can possibly receive after Person X's transfer and Person Z's allocation is 50 points.
B. At the time Person X chooses a transfer to Person Y, Person X does not yet know about the existence of Person Z or about the 40 additional points.
C. When choosing an allocation of the 40 points, Person Z knows which transfer Person X chose.

Question 3
Which of the following statements is true?

A. At the end of the experiment, Person Y's final bonus payment is based onan endowment of 5 adjusted for any transfers chosen by Person X plus any points received from Person Z's allocation.
B. Person Z chooses how much to transfer between Persons X and Y from their initial endowments of 15 and 5.
C. Person X chooses how to allocate 40 additional points between X and Y.

Decision of Person Z

Remember that Person X may:
(the following appears only where negative transfers are possible, where L is the most negative number)

- Transfer 1 to L points **from Y to X**.
- That is, Person X may choose a **negative** transfer to Y of –1 to –L points.

(in some cases, only 1 negative point can be transferred rather than a range of points, in which case above it reads "Transfer 1 point from Y to X")
OR
(the following appears where zero transfers are possible in combination with positive and/or negative transfers.)

- Leave the points of X and Y **unchanged**.
- That is, Person X may choose a transfer to Y of **0** points.

OR

(the following appears only where positive transfers are possible, where H is the highest positive number)

- Transfer 1 to H points **from X to Y**.
- That is, Person X may choose a **positive** transfer to Y of +1 to +H points.

For each possible transfer by Person X, you may try out different values from 0 to 40 to allocate to X from the pull-down menu in the field below marked "Choose." This amount will be added to Person X's points, and the remaining amount from the 40 points will be added to Person Y's points. You may revise any choices at any time before submitting them. Once you are satisfied with all choices, click on "Submit final decisions" at the bottom.

Suppose Person X transfers –L (or 0 or +1) points to Y		**Points after X's transfer** X	Y	**I allocate this amount to Person X**	**Total points** X	Y
−5 *(or –L or 0 or +1)*	20	0		Choose		
−4 *(or whatever)*		19	1	Choose		
−3 *(or whatever)*		18	2	Choose		
(etc.)}						

[<Additional Allocation

Persons Z

In addition to Persons X and Y, other participants have been randomly assigned to a third group consisting of participants called Persons Z. Each Person Z is randomly matched with a Pair consisting of one Person X and one Person Y.

Z Payment

For making a decision about X and Y, each Person Z receives a fixed bonus, which has nothing to do with the eventual payments received by Persons X and Y.

Decision of Person Z

Each Person Z allocates 40 additional points between Persons X and Y. These 40 points will be added to the amounts X and Y receive after the transfer you (Person X) made from the initial endowments, which total 20 points (15 to X and 5 to Y). Person Z is told the amount Person X transferred to Person Y and then allocates an amount out of the 40 additional points to Person X. Any remaining points out of the 40 that Z does not allocate to Person X goes to Person Y, so no points are lost.

Final Payments

The final bonus payments to X and Y are based on their Endowments adjusted for X's transfer plus the points Person Z allocates to them.>]

(Benevolent Dictator Treatments
continues below with Instructions for Person <X> [Y] {Z})

[<Persons Z
In addition to Persons X and Y, other participants have been randomly assigned to a third group consisting of participants called Persons Z. Each Person Z was randomly matched with a Pair consisting of one Person X and one Person Y.

Z Payment
For making a decision about X and Y, each Person Z received a fixed bonus, which had nothing to do with the eventual payments received by Persons X and Y.>]

Decision of Person Z
[<The decision of Person Z was>] {As Person Z, your decision is} to choose how many points to transfer between Persons X and Y. [(As Person)] <X> [Y][<, you make no decision. Specifically, Z may>] {Persons X and Y make no decisions. Specifically, you may}:
(the following appears where negative transfers are possible, where L is the most negative number)

- Transfer 1 to L points **from Y to X**.
- That is, [<Z>] {you} may choose a **negative** transfer to Y of –1 to –L points.

(in some cases, only 1 negative point can be transferred rather than a range of points, in which case above it reads "Transfer 1 point from Y to X")
OR
(the following appears where zero transfers are possible)

- Leave the points of X and Y **unchanged**.
- That is, [<Z>] {you} may choose a transfer to Y of **0** points.

OR
(the following appears where positive transfers are possible, where H is the highest positive number)

- Transfer 1 to H points **from X to Y**.
- That is, [<Z>] {you} may choose a **positive** transfer to Y of +1 to +H points.

When you have understood these instructions, click "Continue" below to proceed to the quiz about these instructions.

Quiz about Instructions
(same as for the Double dictator treatment)

[<Person X's>] {Your} Decision

[<Below you can see a summary of the Endowments, the transfer chosen by Person Z, and the Points after transfer.>]

{Here again are the transfers you may choose for your decision. You may:

- Transfer 1 to L points **from Y to X**.
- That is, you may choose a **negative** transfer to Y of –1 to –L points.

OR

- Leave the points of X and Y **unchanged**.
- That is, you may choose a transfer to Y of **0** points.

OR

- Transfer 1 to H points **from X to Y**.
- That is, you may choose a **positive** transfer to Y of +1 to +H points.

Click on a number in the pull-down menu below in the field marked "Choose" to try out different values for your transfer. Below this, the Points after transfer will update for Persons X and Y. You may revise your choice at any time before submitting it. Once you are satisfied with your choice, click "Submit" below.}

	Person X	Person Y
Endowments	15	5
[<Person Z transferred>] {I transfer}this amount		
Points after transfer		

Questions and Payment

Thank you for your participation in this study. Your payment is
$

To receive this payment, please complete the questions below. Your responses will never be associated with you personally. When you have completed all questions, click "Proceed to get code" below.

1. Age in years

_____

2. Race: please choose the category that best describes your racial/ethnic background
   1 African-American (non-Hispanic)
   2 Asian-American/Pacific-Islander
   3 Caucasian (non-Hispanic)
   4 Hispanic/Latino
   5 Native-American (Indian, Eskimo, Hawaiian)
   6 Mixed Race

3. Gender
   1 Male
   2 Female

4. Marital status
   1 Married
   2 Widowed
   3 Divorced

4 Separated
5 Never married

5. Highest level of education or degree completed
1 Less than high school degree
2 High school degree or equivalent
3 Some college but no degree
4 Associate degree
5 Bachelor degree
6 Graduate degree

6. Employment status
1 Employed, working 40 or more hours per week
2 Employed, working 1-39 hours per week
3 Not employed, looking for work
4 Not employed, NOT looking for work
5 Retired
6 Unable to work

7. Total annual income of all members of your household in US dollars. Please enter without commas:
$_____ per year

(For Z in the Benevolent Dictator treatments and X in the Double Dictator treatments)
{<8. Why did you choose the transfer of points between X and Y as you did?>}

(For Z in the Double Dictator treatments)
{8. Why did you choose to allocate the 40 points as you did?}