2-28-2022

Moral Salience and Conditional Altruism: Reconciling Jekyll and Hyde Paradoxes

James Konow
Loyola Marymount University, jkonow@lmu.edu

FRANK COMMENTS WELCOME (NOW BEFORE YOU REFEREE IT)

February 2022

Moral Salience and Conditional Altruism: Reconciling Jekyll and Hyde Paradoxes

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

The results of many observational and experimental studies reveal an economically and socially important paradox: 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. These patterns are inconsistent with both theories of rational self-interest as well as with theories that incorporate stable social preferences. This paper introduces a theory that reconciles many of these phenomena, including the depressing effects on moral behavior of experimentally introducing uncertainty, social distance, exit options, and possibilities to take from or destroy the earnings of others. The theory combines the concepts of moral salience and conditional altruism to explain not only the paradoxes but also a wide range of classic findings on social preferences.

Keywords: moral salience, conditional altruism, fairness, altruism, moral wiggle room

JEL Classification: C9, D3, D9

Acknowledgements: I wish to thank Zachary Grossman, Prachi Jain, Joel Sobel, Stefan Traub and participants in seminars at Helmut Schmidt University Hamburg and Pomona College and at the North American Economic Science Association for comments and suggestions, and Veronica Backer-Peral and Eamon Shaw for research assistance.

1. Introduction

A pedestrian gives money to beggars but, if possible, crosses the street to avoid them. Different ethnic groups live peacefully together, until genocide normalizes the destruction of life and property, as during the Bosnian War. Otherwise law-abiding citizens join in looting during civil disturbances and natural disasters. Donors in developed countries give to local causes but neglect more critical need in distant developing countries, where their support could do much more good. Some people employ uncertainty about climate change as an excuse not to act on it, even when they support measures to address less severe environmental issues. There are countless economically and socially important instances such as these of "Dr. Jekylls," who under certain circumstances act morally, but then, for reasons that seem morally irrelevant, behave less morally or even immorally, transmogrifying into "Mr. Hydes."

Of course, the examples above do not necessarily require an appeal to inconsistent moral preferences but might instead be explained by a variety of other factors, such as risk preferences, social image concerns, imperfect information, preemptive retaliation, strategic self-interest, fear of punishment, or expectations about the behavior of others. Nevertheless, the results of laboratory and field experiments demonstrate that these paradoxes persist, even when one carefully controls for such factors. In an initial round that began in the 1980s (e.g., Güth, Schmittberger and Schwarze, 1982, Camerer and Thaler, 1995), experimental economists began uncovering instances of behavior at variance with the single-minded pursuit of material self-interest. In time, these initial anomalies became the "classic" results, which prompted numerous theories of stable moral (or social) preferences (e.g., see Camerer, 2003). Then, in the late 2000s, seminal experimental work, including by Dana, Weber and Kuang (2007) on "moral wiggle room," produced new "anomalies." Subjects act fairly under certain conditions but then act unfairly under slightly different conditions in ways that are inconsistent both with pure self-interest and with theories that combine self-interest with stable moral preferences.

One such anomaly, studied by Bardsley (2008) and List (2007), is what I will call the "taking effect." In a version of an experiment called the dictator game, two subjects have initial endowments of money, and one of the subjects, called the dictator, has a larger endowment and may anonymously transfer an amount to the other subject, called the recipient. Most dictator experiments show that most dictators transfer a positive amount. But when dictators are permitted not only to give but also to take from recipients' endowments in an otherwise identical

treatment, many dictators take, and positive transfers also decrease in frequency. For dictators, who are fair enough to share in the standard version, the option to take in the second treatment should not matter, but it often does.

This paper introduces a theory that is consistent with a wide range of stylized facts, including classic findings about social preferences as well as the newer anomalies that contradict both pure self-interest and stable moral preferences. It provides guidance about the conditions under which one can retain a social preference approach and when and how to extend it to account for anomalies. The proposed theory describes the preferences of a decision-maker, called the *agent*, who chooses an action that materially affects a passive person, called the *patient*.[1] One example of the class of decisions considered is a donor's choice about how much to give anonymously to someone supported by a charity, and another example is the aforementioned dictator game. In general, the agent's utility function consists of material utility, which is a function of the agent's own allocation, and of moral preferences over the allocation of the patient. The specific moral preferences addressed here, called conditional altruism, are allocative preferences that consist of two parts. First, fairness captures the disutility experienced by the agent as the patient's payoff falls short of or exceeds the fair amount, similar to inequity aversion terms found in various social preference models. Second, altruism represents the agent's utility from transferring an amount to the patient and disutility from taking an amount from the patient (or, in the case of a spiteful agent, the disutility from giving and utility from taking). Fairness and altruism are weighted by moral salience, which is a function of the decision context, i.e., of the choices and information about the choices. This weight increases with moral context, e.g., opportunities to share, and decreases with non-moral context, e.g., opportunities to take.

The paper "Virtue Preferences: Jekyll and Hyde Paradoxes with Sanctions" (henceforth Konow, 2022a) extends the theory by introducing virtue preferences, which represent a desire to reward or punish another beyond what is called for by fairness alone. It then applies this complete theory to classic findings on reciprocity and to additional anomalies, including outcome bias, willful ignorance, and delegation. It also reports the results of a new experiment that tests the theory out-of-sample and proves supportive of it while providing additional insights into the taking effect. The theory is related to the oldest school of thought in Western moral

[1] Lacking a general and commonly agreed upon term in economics for a person who is acted upon by a moral agent, I borrow the term patient from philosophical ethics.

philosophy, virtue ethics. Referring 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 theory presented here and in Konow (2022a) relates to other features of virtue ethics, including multiple ethical principles, context-dependent morality, and a role for preferences over virtues.

Of course, strategic interactions are extremely important, but the theory is formulated around explaining and predicting behavior in simple non-strategic experiments because of several advantages of that approach for the task at hand. A growing literature has demonstrated the external validity of non-strategic experiments for moral preferences quite generally, that is, pro-sociality in experiments is correlated with important behaviors in the field. For example, dictator generosity is positively correlated with honesty in the field (Franzen and Pointner, 2013) and with a willingness to take costly steps to reduce the exposure of others to Covid-19 (Campos-Mercade, Meier, Schneider and Wengström, 2021). In addition, the more recent dictator-style experiments on anomalies provide persuasive evidence of the internal validity of the claim that there is something inherent to moral preferences that is inconsistent with existing theories. Moral preferences are clearly relevant to important economic phenomena, such as cooperation, but cooperation is impacted by a complex set of considerations, as Dal Bó and Frechette (2018) argue. Specifically, strategic self-interest can confound inferences about morals in many contexts but should play no role in non-strategic decisions such as the dictator game. In particular, "virtue signaling," or feigning morally motivated behavior for strategic reasons, can distort signals about true moral preferences, which is another reason for the focus here on non-strategic decisions. Finally, simple experimental decisions enable the parallel development of a simple and tractable theory. That said, reference will occasionally be made to results where strategic concerns play a potential role, in particular, where results from non-strategic designs are unavailable but experiments exist for which strategic concerns are likely negligible.

A word is in order about what this paper tries, and does not try, to do. It proposes a theory of moral preferences that is novel, tractable, and capable of explaining a wide range of evidence on moral preferences, including various Jekyll and Hyde paradoxes. It makes some comparisons with alternative explanations, but, for various reasons, it does not conduct a beauty contest among theories. Similarly, it focuses on the consistency of the many theoretical predictions with numerous stylized facts from previous studies rather than on statistical tests of the findings of

those studies, which were not designed to test the theory (statistical analysis of an experiment specifically designed to test this theory can be found in Konow, 2022a). Moreover, with respect to its chief theoretical ambitions, I find various theories plausible in the particular cases they address, so the aim here is not to displace them. Instead, I see this paper as offering a theoretical framework that is new, distinct from others, and tractable. Moreover, it is comparatively specific in its predictions and general in its applications to many types of behavior that are impacted by moral preferences, including paradoxes that have resisted similarly general explanations, which it traces to a common cause. It might be seen as a generalization of prior theories.

Section 2 introduces the theory of moral salience and conditional altruism. The next five sections apply the theory to explain effects on generosity of proximity (section 3), as well as more unexpected effects of uncertainty (4), taking options (5), opportunities to destroy the wealth of others (6), and possibilities to avoid situations of giving (7). Section 8 demonstrates the consistency of the theory with numerous classic results on moral preferences, and section 9 concludes.

2. Theory

This section introduces moral salience, generalizes the model of conditional altruism, and merges the two. In neuroscience and social psychology, salience refers to how an object stands out relative to its environment, and economists have cited salience in a similar sense. For example, labeling a choice in a salient manner can increase coordination in experiments, e.g., Crawford, Gneezy and Rottenstreich (2008) and Crawford and Iriberri (2007). Chetty, Looney, and Kroft (2009) find that posting prices inclusive of taxes reduces demand based on both experimental and observational data, which the authors attribute to the salience of posted prices. Bénabou and Tirole include salience variables in their theoretical treatments of moral identity (2011) and beliefs about fairness (2006). To my knowledge, Bordalo, Gennaioli, and Shleifer are the first in economics to formulate salience as dependent, rather than solely an independent, variable. They invoke salience to explain the endowment effect (2012) and anomalous consumer choices (2013, 2016), and they focus on the salience of attributes of an individual good (e.g., its price and quality). The *moral salience* proposed here differs in several respects from their or any other prior formalizations, to my knowledge. I introduce a theory of *set salience*, which proposes a simple but novel function that characterizes how properties of subsets, rather than individual elements, of a set affect the prominence of the subsets. Moreover, set salience is applied to moral

contexts, viz., to how morally good and morally bad contexts affect moral preferences and choices. Below I begin with a description of the decision context, proceed to measures of good and bad contexts, and then specify moral salience.

Consider an agent, who makes a decision that materially affects a passive patient. This might be a sponsor choosing how much to donate to a child supported by a charitable organization or a dictator deciding how much of an endowment to transfer to a recipient in a dictator game. The agent may take an action, *x*, from the set of available actions, *X*. In the situations considered in this paper, the action is the same as the material effect on the patient, e.g., the transfer received by the recipient in a dictator game, which is selected from the set of permissible transfers, X. The agent also possesses information that the agent might see as morally relevant to the choice of actions. For example, a dictator could be informed that the recipient is socially distant, *y*, among other elements of the information set, Y. Indeed, mere labels or even false information might be relevant. Actions and information are disjoint proper subsets of the decision context, C, i.e., $X \cap Y = \emptyset$, and $C = X \cup Y$, which is, in the interesting case, non-empty.

Moral salience is the weight attached to the agent's moral preferences as a result of the decision context. For example, the taking effect described in the Introduction is consistent with the interpretation that adding taking options to the set of available actions in a dictator game reduces the weight on moral preferences through X and, therefore, the level of dictator transfers. The same effect on moral preferences and transfers might be achieved through Y, e.g., by characterizing a dictator's task as an "exchange" rather than a "division" or by underscoring the dictator's anonymity (Hoffman et al., 1994, 1996). We will consider in future sections various ways in which qualitative elements of context can affect moral salience, but for the discussion below it is helpful to think of quantitative contextual elements $c_i \in \mathbb{R}$, e.g., giving and taking options in a dictator game or the physical distance from the dictator to the recipient.

The decision context C can be partitioned in an additional manner according to the effects of those partitions on moral salience. Moral context, denoted C_+, increases moral salience, e.g., opportunities to help another person, such as amounts a dictator may give to a recipient. Non-moral context, denoted C_-, decreases moral salience, e.g., opportunities to harm another, such as amounts a dictator may take from a recipient. Individual elements of both X and Y might be categorized as moral or non-moral, where $C = C_+ \cup C_-$ and $C_+ \cap C_- = \emptyset$. One might further

distinguish amoral, or morally neutral, elements, e.g., the possibility of inaction, such as neither giving nor taking, although for the cases considered here, this can be folded into moral context.

Moral salience itself is based on measures of moral and non-moral context. Define a function, $m(C_i)$, of partitions, C_i, of the context that satisfies the properties of a measure, viz., non-negativity ($m(C_i) \geq 0 \forall C_i$), null empty set ($m(\emptyset) = 0$), and countable additivity ($m(\bigcup_i C_i) = \sum_i m(C_i)$). Specifically, let us partition the context into its moral and non-moral subsets, i.e., $C_i = \{C_+, C_-\}$, and denote the moral measure $p \equiv m(C_+)$ and the non-moral measure $n \equiv m(C_-)$. The moral measure is increasing in moral context, and the non-moral measure is increasing in non-moral context. Distinguishing moral and non-moral context and constructing measures of them requires, of course, some judgment and depends on the decision context. In the various applications that follow, we discuss different commonsensical specifications for these measures.

Now we come to moral salience, which is related to the usual understanding of salience in neuroscience and social psychology, where salience typically refers to how an object stands out relative to its environment. Here, I propose a novel specification of salience that I call "set salience," which involves collections of objects that are all disjoint subsets of a superset. I focus on the case in which the context may be bifurcated into measurable subsets. Set salience refers to the tendency for the subset with smaller measure to have disproportionate prominence relative to the contrasting subset with larger measure. For example, a five-year-old does not stand out in a Kindergarten but does in a retirement home. The type of salience introduced here further specifies a non-linear relationship between salience and measures of the subsets of context. For example, a comparatively small group of children situated among older people is prominent, but the marginal salience of the first child is greater than that of the second and the marginal salience of the second is greater than that the third, etc.

Moral salience is an application of set salience to moral contexts. It describes how subsets of elements of the decision context affect the prominence of moral considerations and, therefore, the weight on an agent's moral preferences. The elements of each subset share some feature(s), here, whether they are moral or non-moral, and each subset distinguishes itself in this way from the other. Moral salience formalizes this property for moral preferences. Analogous to the anthropomorphic example above, the addition of elements of non-moral context to a given moral context, and the attendant increase in the non-moral measure, decreases moral salience at a

decreasing rate, that is, the first addition non-moral context causes a larger decrease in the prominence of moral considerations than the second, etc. Conversely, the addition of moral context to a given non-moral context, and the related increase in the moral measure, increases moral salience and does so at a decreasing rate. Formally, consider the following definition, which reflects these properties.

DEFINITION 1: Moral salience, $\sigma(p,n)$, is a function that maps the moral and non-moral measures of the decision context into the half-open unit interval:

$$\sigma\colon \mathbb{R}_+^2 \to (0,1].$$

It is assumed that $\sigma(p,0) > 0, p > 0;\ \sigma(0,n) \geq 0, n > 0$; and that $\sigma(p,n)$ is twice continuously differentiable with

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

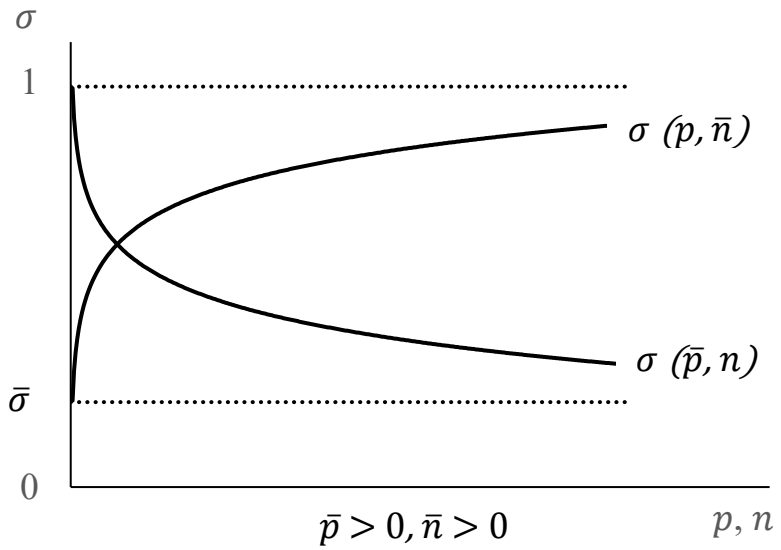
It proves convenient in the subsequent analysis to flesh out this function in a more specific form. One expression that captures the assumed relationships is the ratio $\frac{p}{p+n}$, which is defined, since I assume throughout that $p > 0$. Many decisions, though, involve some fixed moral salience with variation in only a subset of the moral context. For example, in a dictator game, variation in the amounts one may transfer might impact moral salience through its effects on *p* or *n*, but there are often some baseline moral considerations, e.g., triggered by the very fact of being endowed and paired with another person. In such cases, the context contains a baseline, or fixed, moral set salience denoted $\bar{\sigma} \in [0,1)$ in addition to the subsets of moral context that are variable, *p* and *n*. This leads to the following specification for moral salience:

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

In the cases considered here, I assume $\bar{\sigma} > 0$ and/or $p > 0$. Under these assumptions, this expression satisfies the conditions that define moral salience: $\sigma > 0$, its maximum is $\sigma(p,0) = (1-\bar{\sigma}) \cdot \frac{p}{p} + \bar{\sigma} = 1$, its minimum is $\sigma(0,n) = \frac{0}{n} = 0$ when $\bar{\sigma} = 0$, $\frac{\partial\sigma}{\partial p} = (1-\bar{\sigma})\frac{n}{(p+n)^2} > 0$, $\frac{\partial^2\sigma}{\partial p^2} = -(1-\bar{\sigma})\frac{2n}{(p+n)^3} < 0$, $\frac{\partial\sigma}{\partial n} = -(1-\bar{\sigma})\frac{p}{(p+n)^2} < 0$, and $\frac{\partial^2\sigma}{\partial n^2} = (1-\bar{\sigma})\frac{2p}{(p+n)^2} > 0$.

Figure 1 illustrates how moral salience varies with the moral and non-moral measures. The aforementioned effects of moral and non-moral context on moral salience are reflected in the properties of σ being increasing and concave in *p* for a given $\bar{n} > 0$ and decreasing and convex in *n* for a given $\bar{p} > 0$. Note that $\bar{\sigma}$ is the greatest lower bound of moral salience.

The remaining sections analyze numerous contextual factors that affect moral salience. Some cases involve binary decisions, such as whether or not to remain ignorant of information that raises a moral obligation. In other cases, however, there is empirical evidence on the effects of incremental changes in moral or non-moral context. For instance, moral salience may vary with the amounts that may be given to or taken from a patient, physical proximity to the patient, and probability that the agent's decision is actualized. These cases lend themselves to cardinal measurement, so that one can observe not only the direction of the effect of context on moral



salience but also differences in the rate of change in that effect.

FIGURE 1. – Moral salience.

Another practical aspect is that people are sometimes confronted with multiple decisions in similar moral contexts at the same time. This occurs, for example, in experiments that present the same group of subjects with similar decisions in a within-subjects design. It also arises, though, outside the laboratory, e.g., when someone receives multiple solicitations to donate to different charities. In such cases, I make the following assumption.

ASSUMPTION 1: Let an agent make decisions 1 and 2 in contexts C^1 and C^2, respectively. Suppose 1 and 2 are related, meaning choices are made jointly from decision contexts that are identical except for some element, c: $\{C^1/c^1\} = \{C^2/c^2\}$ and $c^1 \neq c^2$. Then the decisions share the common context $C = \{C^1 \cup C^2\}$ with the same moral salience and same measures, p and n.

Finally, let us clarify the kinds of contextual factors that are assumed to affect moral salience and how and do so in both general terms as well as through some applications.

ASSUMPTION 2: Moral salience is a decreasing function of factors that increase the perceived separation between the agent's choice and the moral consequences of that choice on a patient. Factors that can increase perceived moral separation (and, thereby, reduce moral salience) include increases in the set of harming choices (e.g., taking or destroying the patient's wealth), the degree of uncertainty about the consequences of the choice, and opportunities to avoid the choice, as well as decreases in the perceived membership of the patient to a group to which the agent is morally responsible.

These concepts will be fleshed out in greater detail in the applications of the theory to different classic and anomalous results in the following sections of the paper. To reinforce an earlier point, note that, since context includes the information provided, *Y*, moral salience can be subject to framing effects, e.g., a dictator's transfer can be affected by labeling such as whether the task is worded as a choice to "give" or to "distribute" money.[2] But the effects on moral salience are not limited to framing effects, since moral salience is also a function of the actual set of available choices, *X*, and not just their presentation, e.g., whether a dictator may take as well as give. Thus, context can affect choices mediated by salience even under perfect information due to differences in choice sets.

As already stated, the theory presented here weights moral preferences by moral salience, but, given the rich set of moral preferences, a critical question concerns which moral preferences. For concreteness and tractability, this study focuses on allocative preferences, specifically, those addressed by the model of *conditional altruism* introduced in Konow (2010). The following discussion extends this model, generalizing the altruism term, elaborates new implications of the model, and analyzes some implications of integrating moral salience into it.

Conditional altruism has three components: material utility and two moral motives, fairness and altruism. Material utility, $u: \mathbb{R}_+ \rightarrow \mathbb{R}_+$, is assumed to be a twice continuously differentiable function of the agent's material allocation, π_a. Specifically, I assume material utility, $u(\pi_a)$, is as follows

$$u(0) = 0, {}^{\partial u}/_{\partial \pi_a} > 0, \text{ and } {}^{\partial^2 u}/_{\partial {\pi_a}^2} \leq 0.$$

The "conditional" part of conditional altruism involves fairness, which is conditioned on

[2] See Bergh and Wichardt (2018) for evidence of the effects of this wording on dictator transfers. In fact, information might even be false but relevant to the agent's actions, if it influences moral salience, although providing false information is typically taboo in economics experiments.

the fair allocation to the patient or so-called *entitlement*. Fairness is a type of inequity aversion, i.e., it captures the disutility experienced by the agent as the patient's allocation differs from the entitlement. The entitlement refers here quite generally to distributive norms, and this is another avenue through which context works its way into the theory. Moreover, it affects the allocative decisions of *stakeholders*, such as dictators in dictator games, as well as of impartial third-party allocators, or *spectators*. I make the following assumption about the entitlement.

ASSUMPTION 3: The entitlement, denoted η, is the moral allocation by the agent to the patient. It depends on the salient distributive norm or norms in a given context, *C*, which can be inferred from the allocations of spectators in *C*. In "simple" contexts where no specific norms are salient, the entitlement reduces to equality.

A large literature on stakeholder and spectator decisions has now established the effects of multiple distributive norms. The use of spectators to elicit moral norms was introduced in Konow (2000) and has been employed and elaborated in numerous subsequent studies, including Aguiar, Becker and Miller (2013), Almås, Cappelen and Tungodden (2020), Cappelen, Konow, Sørensen and Tungodden (2013), Croson and Konow (2009), Konow (2012), Konow, Saijo and Akai (2020), and Møllerstrom, Reme and Sørensen (2015). Distributive norms can also be inferred from stakeholder decisions, e.g., Cappelen et al. (2007), and the norms based on stakeholder decisions have been shown to be equivalent to those inferred from spectators, e.g., Cappelen et al. (2013a). Together these studies have demonstrated the dependence of distributive norms on the context, e.g., efficiency is relevant in contexts where total surplus is variable, e.g., Møllerstrom et al. (2015), equity (i.e., proportionality) in contexts where people choose different contributions to surplus, e.g., Konow et al. (2020), need in contexts with information about basic needs, e.g., Konow (2010), and equality in simple contexts that are low in morally relevant information about differences among individuals, e.g., Croson and Konow (2009, RZ treatment) and Konow (2000, benevolent/exogenous treatment).

Although the entitlement may be derived from stakeholders, Assumption 3 singles out spectators for several reasons. Most importantly, spectators provide a measure of the entitlement that is independent of stakeholders and is robust to very general assumptions about the functional form of inequity aversion. It avoids the need, otherwise, with stakeholders for a priori assumptions about the form of inequity aversion or for a posteriori assumptions that might be accused of selection for their fit to the particular results. Spectator decisions also provide a less

dispersed measure, since they are not distorted by stakeholder self-interest that differs across agents. Those very stakeholder interests can also produce biased beliefs about the relevant norm. Although bias is potentially relevant to stakeholder decisions, the theoretical treatment here requires only that biased and impartial beliefs co-vary directly with context, and evidence from studies of both spectators and stakeholders indicate that they do, e.g., Konow (2000) and Konow et al. (2020). Finally, although multiple entitlements are relevant to certain findings discussed in the expanded theory in Konow (2022a), the experiments considered here can be reconciled with a single entitlement, which in most experiments reduces to equality.

Fairness is expressed as a function, $f: \mathbb{R} \to \mathbb{R}_{\leq 0}$, that captures the preference of the agent over the material allocation of the patient, π_p, relative to the patient's entitlement, η. Specifically, I assume f is the twice continuously differentiable function

$$f(\pi_p - \eta),$$

where $f(0) = 0$, $\partial f / \partial w \cdot w < 0$ for $w \equiv \pi_p - \eta \neq 0$, and $\partial^2 f / \partial w^2 < 0$. Agents are assumed to differ in the strength of their fairness preference, which is captured by the fairness coefficient $\phi \in \mathbb{R}_+$ that is applied to f to form $\phi f(\pi_p - \eta)$. This coefficient is distributed according to the cumulative distribution function $\Phi(\phi)$, where $\Phi(\phi)$ has support $[\underline{\phi}, \overline{\phi}]$ with $0 < \underline{\phi} < \overline{\phi} < \infty$ and $0 < \Phi\left(\underline{\phi}\right) < 0.5$. The assumptions about $\underline{\phi}$ help establish predictions that are consistent with behavior discussed later, viz., that all agents care somewhat about fairness and that minimally fair types constitute a minority. Note that agents experience disutility, when patients have more or less than their entitlement. That is, fairness is never utility increasing, which reflects the idea that moral norms signify an obligation rather than an opportunity. This is a critical factor later for explaining a number of empirical findings.

The "altruism" part of conditional altruism refers to a moral preference that is personal and unconditional. As with standard theories of altruism, it is not conditioned on a moral norm, such as equity or efficiency, or on the behavior of others, such as a desire to reward or punish deviations from norms. Here I generalize the prior version of this model, which formally resembled warm glow (e.g., Andreoni, 1989), to encompass not only giving but also taking and to include explicitly not only positive but also negative altruism (i.e., spite). Unlike pure altruism but like warm glow, it is assumed to be 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 making a transfer to or

from a recipient in a dictator game. Unlike pure altruism, it is not a function of the patient's total allocation or of any amounts the patient receives from others. Altruism is also personal in that it assumed to apply to agent-patient relationships but not to impartial third party, or spectator, decisions. Altruism is expressed as a function, $g: \mathbb{R}^2 \to \mathbb{R}$, of the amount, x, the patient receives from the agent and the altruism coefficient, α.

Specifically, I assume altruism is the twice continuously differentiable function $g(x, \alpha)$

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 ${}^{\partial^2 g}/_{\partial x \partial \alpha} > 0$. Agents differ according to their altruism coefficient, $\alpha \in \mathbb{R}$, and are categorized as altruistic, $\alpha > 0$, selfish, $\alpha = 0$, or spiteful, $\alpha < 0$. The altruism coefficient is distributed according to the cumulative distribution function $\mathrm{A}(\alpha)$, which has support $[\underline{\alpha}, \overline{\alpha}]$ with $-\infty < \underline{\alpha} < 0 < \overline{\alpha} < \infty$. I assume $\mathrm{A}(0) < 0.5 < \mathrm{A}(\overline{\alpha}) - \mathrm{A}(0)$ and $\int_{\underline{\alpha}}^{\overline{\alpha}} \alpha \, \rho(\alpha) d\alpha > 0$, where $\rho(\alpha)$ is the probability density function of α. That is, I assume a minority of agents is spiteful, a majority is altruistic, and the average type is altruistic. The altruism term accommodates positive transfers to the patient, $x > 0$, as well as negative ones, $x < 0$, i.e., taking from the patient. The giving case is similar to warm glow, but this term additionally incorporates disutility from taking. Note that this term is upward sloping for $\alpha > 0$ and downward sloping for $\alpha < 0$, i.e., the utility of a spiteful agent decreases with giving and rises with taking.

I assume additively separable utility, keeping with most social preference models, e.g., Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Fehr and Schmidt (1999), and Rabin (1993). Letting the moral preference terms be weighted by moral salience, σ, the utility of the agent, U, becomes

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

There are sometimes arguments for separate moral salience variables for different moral preferences, i.e., for distinguishing fairness salience, σ_f, from altruism salience, σ_g. Nevertheless, none of the results discussed here depends on such independent variation, so I simplify the analysis and use a single moral salience term, σ. Finally, when uncertainty is involved, I assume that decision-makers are expected utility maximizers.

Since the focus of much of the analysis is on the dictator game, we will interpret equation

(2) for the general case of this standard game. Specifically, let *X* represent the endowment of the dictator and *x* the amount the recipient receives as a consequence of the dictator's transfer such that $\pi_a = X - x$. The recipient's endowment, in versions of the dictator game where it is relevant, is denoted *Y* such that $\pi_p = Y + x$. Then, for the general dictator game, we have the following equation:

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

I add one more assumption to accommodate a stylized fact of standard dictator games: a minority of dictators makes super-fair, i.e., larger than fair, transfers. Let σ^* denote the level of salience in the standard dictator game and α^* the value of α in that game such that marginal altruism equals marginal material utility when evaluated at the fair transfer, i.e., $\alpha^* = \{\alpha | \, {\partial u}/{\partial \pi_a}\,(X - \eta) = \alpha^* \cdot {\partial g}/{\partial x}\,(\eta, \alpha)\}$. Then I assume $0 < \alpha^* < \overline{\alpha}$ and $0 < \mathrm{A}(\overline{\alpha}) - \mathrm{A}(\alpha^*) < 0.5$. This implies that a minority of dictators in the standard game is so altruistic that their utility maximizing transfer is greater than the fair amount.

Now we turn to some theorems about transfers in the dictator game, which will come in handy in the later analysis. The proofs appear in the Appendix. I begin with the effects of moral salience.

THEOREM 2.1: The optimal transfer, *x*, is increasing in σ.

This is due to the increased weight on moral preferences. As stated above, we proceed from high moral salience and focus on the effects of non-moral context, *n*, on reducing salience. So, it proves useful to establish the relationship between *n* and x, which is addressed in Theorem 2.2.

THEOREM 2.2: The optimal transfer, *x*, is decreasing in *n*. Assuming *x* is weakly convex in σ, *x* is strictly convex in *n*.

This theorem states that non-moral context decreases giving due to the reduction in moral salience. In addition, it establishes that, in the case of cardinal measures of *n*, giving decreases at a decreasing rate due to the strict convexity of σ in *n*, assuming *x* is weakly convex in σ.[3] That

[3] Note that the assumed relationship between *x* and σ is a feature of several commonly used parametric utility functions. For example, for the standard dictator game, suppose we can write $U = X - x + \sigma h$, where $h = f + g$. This formulation treats material utility as linear in the dictator's payoff, as commonly assumed in many social preference theories, e.g., Charness and Rabin, 2002, Dufwenberg and Kirchsteiger, 2004, Falk and Fischbacher, 2006, Fehr and Schmidt, 1999, and Rabin, 1993, and as a special case of the current theory (since $u' > 0$ and $u'' \leq 0$). This assumption seems innocuous for economics experiments, where the stakes are usually modest relative to subjects' overall income or wealth. Then, it is straightforward to show that, if $h = a + b \cdot \ln x\,, a, b > 0$, then

is, the initial addition of non-moral context causes a larger decrease in giving than the next.

For two reasons, the remainder of this paper focuses on the effects of variation in n rather than p. First, theory yields a stronger prediction about n: from Theorem 2.2, $\frac{\partial^2 x}{\partial n^2} > 0$, but the sign of $\frac{\partial^2 x}{\partial p^2}$ is ambiguous due to the concavity of σ in p. Second, most of the experimental evidence related to moral salience is related to variation in n. Moreover, the few studies of which I am aware that explicitly relate to variation in p are consistent with the theoretical claim that $\frac{\partial x}{\partial p} > 0$ but do not involve cardinal measures needed to shed light on the second derivative, e.g., dictator giving increases significantly with the addition of a short statement about the recipient's reliance on the dictator (Brañas-Garza, 2007), and trading volume in an experimental market decreases when a negative externality is added (Sutter et al., 2020).

Finally, consider the effects on transfers of changes in ϕ, α, and η.

THEOREM 2.3: The optimal transfer is increasing in the fairness coefficient, ϕ, except for super-fair dictators, for whom it is decreasing in ϕ.

THEOREM 2.4: The optimal transfer is increasing in α.

THEOREM 2.5: The optimal transfer is increasing in the entitlement, viz., $0 < {}^{dx}/_{d\eta} < 1$.

Thus, stronger moral preferences generally result in higher transfers. An exception is explicitly noted for super-fair dictators, who experience increased disadvantageous inequity aversion. Theorem 2.5 implies that a one unit increase in the entitlement produces a less than one unit increase in the transfer, which is due to diminishing marginal altruism and the increasing marginal material disutility.

Now we turn to applications of the theory to numerous stylized facts (abbreviated SF), including about mean behavior of agents or subgroups of agents, and the distribution of behavior based on salience, fairness types, altruism types, and action sets.[4] Some stylized facts are ancillary and simply taken as empirical regularities without proof. But the main SFs of interest are numbered and accompanied by (and sometimes, as in the first case below, identical to) a

---

${}^{dx}/_{d\sigma} = b > 0$ and ${}^{d^2x}/_{d\sigma^2} = 0$, or, if $h = a + b \cdot x^{1/2}, a, b > 0$, then ${}^{dx}/_{d\sigma} = \frac{1}{2}b^2\sigma > 0$ and ${}^{d^2x}/_{d\sigma^2} = \frac{1}{2}b^2 > 0$, both of which are consistent with (weak) convexity of x in σ, and, therefore, strict convexity of x in n.

[4] One type of stylized fact is not treated here, viz., preference-based masses that are observed in many social preference experiments, e.g., a spike at equal splits in many standard dictator games, but Konow (2022a) provides an explanation based on different type of moral salience.

theorem that asserts a claim about the consistency of the SF with the theory.

3. Moral Proximity

The Make-A-Wish Foundation is a non-profit organization founded in the United States that uses the contributions of donors to fulfill the wishes of children with life-threatening illnesses. Although stories of these children being granted their wishes tugs at the heartstrings of any compassionate person, the philosopher Peter Singer maintains that the resources of American donors would be put to better use helping people in developing countries (2013). He and other advocates of the philosophical and social movement called "effective altruism" argue for directing charitable resources to where they will do the most good. Singer points out that the average cost in the US of a "wish" (now in excess of $10,000) could, in developing countries, save the lives of at least two or three children if spent on malaria nets or protect 100 children from blindness. Nevertheless, many donors in developed countries favor local or domestic charities over those operating in developing countries.[5]

These conflicting moral intuitions are but one example of an important and common type of anomaly, which, I argue, can be explained by moral salience. This section addresses what I will call moral proximity. This concerns one of the most important practical moral questions, viz., the identity of one's *moral group* or the set of persons to which one feels obligated to be moral. Moral proximity provides an explanation for the effects on moral behavior of, inter alia, physical distance, familial relations, friendship, homophily including by ethnicity or political affiliation, or information about the agent or patient. These effects often seem so intuitive that they hardly strike us as anomalous, although it is still sometimes surprising how easily they can be triggered. And yet they are not predicted by most social preference theories, and I am unaware of theoretical accounts that are able to cast all the different examples in a unified framework.

Most of philosophical ethics concerns moral principles in general terms and scarcely addresses the question of moral groups (although there are exceptions, e.g., Walzer, 1983). Equal moral consideration of all seems noble, but it is self-evident that moral obligations cannot extend indefinitely, and the boundaries are very much in dispute: some people draw the line at family, clan, religious, political or ethnic group, some claim we are obliged to our fellow citizens, some

[5] Singer mentions additional (not mutually exclusive) explanations for the appeal of the Make-A-Wish Foundation, such as the so-called identifiable victim effect. I analyze this effect and its possible causes, including moral salience, in Konow (2022b).

to the unborn, some to all people in the world, and some also to animals (which raises the further question of "which animals?"). Even a broad conception of moral group cannot plausibly maintain that all members of that group are equal: surely, the obligation to one's child differs from that to a securities trader in a distant country. The identification of the moral group is paramount to economic policy. For example, suppose one seeks to promote fair earnings (or efficient earnings or any other normative goal). The first order of business is to identify the set of persons whose earners should be targeted: those within a firm, city, county, state, country, the world? There is also the sticky question of which generations to include, which is critical for so many policies including climate change, i.e., do we include only the current generation or also future ones, and, if so, which? There are practical reasons for favoring one answer or the other, but the moral question must still be factored in, and its resolution is less than obvious.

I will not attempt to resolve these normative questions here, but they are offered as motivation for the importance of the topic and as inspiration for the current descriptive undertaking. The economic importance of the topic is suggested by various phenomena, including by the effects of co-workers on productivity, e.g., Bandiera, Barankay, and Rasul (2010). The focus here is on analyzing of how moral behavior is affected by contextual factors that make the patient's membership in the agent's moral group more or less salient. That is what is meant by moral proximity, and we proceed, as usual when operationalizing moral salience, from a high salience reference point to help identify the properties that affect salience. Assumption 2 in section 2 outlined factors that affect moral salience. To elaborate the proximity aspect of that assumption, consider the following assumption.

ASSUMPTION 4: Patients are most morally proximate and, therefore, salient, when, ceteris paribus, they are physically near, personal information about them is abundant and/or stresses their membership in the agent's moral group, they are associated with other proximate persons, others possess abundant personal information about the agent, agent and patient communicate with one another, and agent and patient share traits in common, even ones that might seem superficial and morally irrelevant. In the case of cardinal measures of distance, let the patient, who is most proximate to the agent, have positive measure, $p > 0$ and the additional distance to a more distant patient be non-moral context, n.

Many of the factors that influence moral proximity have often been characterized as "social distance" (in the pre-Covid-19 sense of the term). For example, transfers rise, when even

limited information about the dictator is provided to the experimenter (Hoffman, McCabe and Smith, 1996), the recipient (Bohnet and Frey, 1999, Grossman, 2015), or both (Alevy, Jeffries, and Lu, 2014). In fact, giving rises, even if the mere existence of a dictator, who remains anonymous, is revealed to the recipient (Dana, Cain and Dawes, 2006). Even three dots on a screen in a "watching eyes" position, instead of a neutral position, can increase dictator transfers (Rigdon et al., 2009). Some of these effects can plausibly be attributed, at least in part, to social image concerns, even under anonymity (see the discussion in the following section of Andreoni and Bernheim, 2009). Nevertheless, social image does not easily explain other factors that fall under the rubric of moral proximity. Dictator giving rises, if the recipient reveals one-way his/her identity to the dictator (Bohnet and Frey, 1999), indeed, even if only the recipient's family name is revealed (Charness and Gneezy, 2008). Conversely, dictator gifts fall, if it is revealed that the recipient is a member of an out-group (Whitt and Wilson, 2007), and Candelo, Eckel and Johnson (2018) report that dictator transfers to a family member are greater than those to a community group or stranger. In addition, transfers increase, if recipients can send a message to the dictator (Bohnet and Frey, 1999, Ellingsen and Johannesson, 2008, Xiao and Houser, 2009).

The variables listed above are not presented as exhaustive, since the question of what affects perceptions of moral groups is an empirical one. Indeed, the plausible examples of moral proximity are too numerous to summarize concisely, so let us focus on studies of physical distance and moral behavior, which are limited in number. In addition, unlike other social distance factors, physical distance is a cardinal measure that permits not only examination of the decreasing effect of non-moral context on transfers but also of the predicted decreasing rate of change. Specifically, consider an agent, who may transfer something of material value to a patient, whereby the distance to different patients varies. The factor that varies in this context, C_i, is physical distance, and $m(C_i)$ is a measure of it. Then, define p and n for this cardinal measure as in Assumption 4. Further, denote the total distance $\delta = p + n$ and normalize the measure of the distance of the most proximate patient, i.e., $p = 1$. Then, remembering our specification for moral salience, this can be expressed

$$\sigma = (1 - \bar{\sigma}) \cdot \frac{p}{p+n} + \bar{\sigma} = (1 - \bar{\sigma}) \cdot \frac{1}{\delta} + \bar{\sigma}, \tag{4}$$

which, if $\bar{\sigma} = 0$, reduces to $\sigma = \frac{1}{\delta}$. It is interesting to note that $\frac{1}{\delta}$ also captures a feature of visual salience, viz., the relationship between distance and perceived size: an object at twice the

distance appears half as large. Consider now recent evidence on giving and physical distance.

SF/THEOREM 3.1: Agent contributions decrease at a decreasing rate with physical distance to patients (Touré-Tillery and Fishbach, 2017, Dejean, 2020, Kühl and Szech, 2017).

PROOF: This follows from Assumption 4 and Theorem 2.2 under the assumptions stated there.

Some recent studies find that physical distance influences contributions to others. Touré-Tillery and Fishbach (2017) report that alumni giving to a large private US university is inversely related to physical distance (Study 2). Figure 2 summarizes the relationships between generosity and non-moral context for six studies. I will return to panels (c) to (f) the next two sections, but note now the consistency of the results with Theorem 2.2 across diverse measures of generosity and diverse measures of non-moral context: they are all inversely related, and the generosity measures appear convex in non-moral context, n, in every case where the data allow its detection, that is, wherever there are more than two levels of n (i.e., except for d). Panel (a) of this figure illustrates the results of a regression based on the data of Touré-Tillery and Fishbach that employs the natural log of distance, which provides a better fit than a linear specification or a non-linear one that adds the square of distance. These results are consistent with moral proximity: donations decrease with physical distance at a decreasing rate. The authors report that the inverse relationship is robust to various controls, including age, income, graduation year, etc.

That said, observational studies cannot rule out omitted variable bias, although they can sometimes shed light on it. Dejean (2020) studies the relationship between rewards-based crowdfunding and physical distance. Using a log specification, he finds investments decrease with distance at a decreasing rate, viz., they are half as large at twice the distance. Nevertheless, the effect of distance is significantly reduced when social networks are taken into account. Social networks are consistent with a different kind of moral proximity, but this effect weakens claims about physical distance, per se.[6] Such issues are not a concern with the experimental studies of anonymous giving by Kühl and Szech (2017), which permit stronger causal inferences about the effect of physical distance. Their field experiment finds that, holding other factors constant, donations to local refugees decrease significantly with distance to their camp, which is varied at two levels. Their laboratory experiment varies distance at more levels and comes to similar

[6] One should be cautious, though, about trying to transfer lessons from Dejean's study to the topic of generosity. Rewards-based crowdfunding arguably relates partially to generosity, given that the rewards are typically uncertain and not commensurate with the investments, but there is often an expectation of some "reward," even if only a thank you note. In addition, the dependent variable in his study is the number of contributions rather than their value.

conclusions, although the relationship is insignificant at longer distances.[7] Figure 2b shows average contributions for their laboratory experiment. Although the convexity appears subtle in this case, that is consistent with the short distances, and the change in the slope is actually similar to that in Dejean and greater than that in Touré-Tillery and Fishbach for comparable distances.

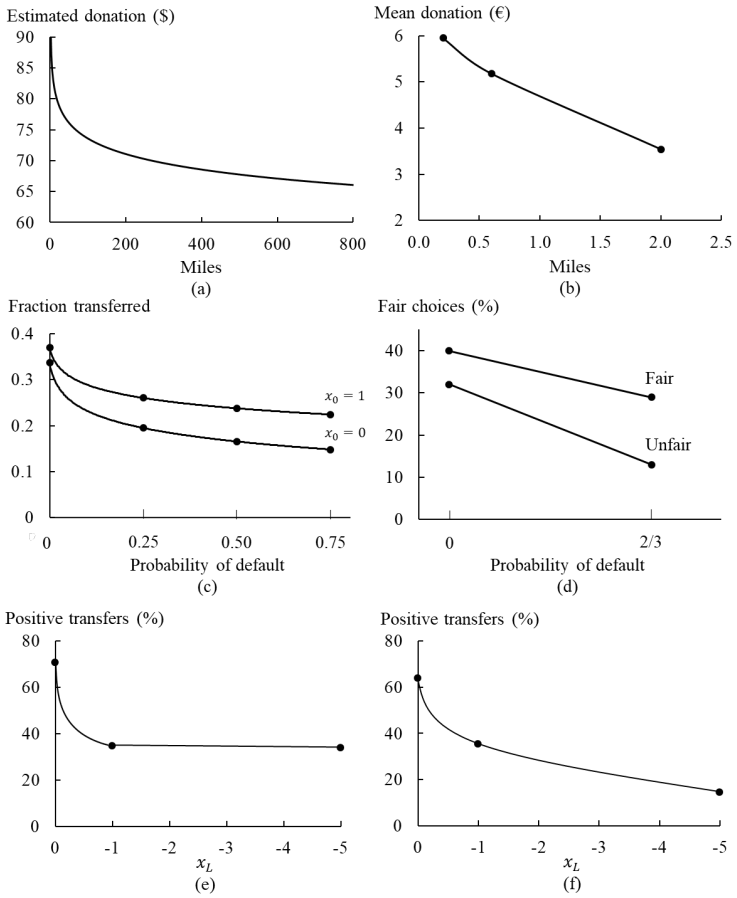


FIGURE 2. – Generosity and Non-moral Context.

Sources: (a) Touré-Tillery and Fishbach (2017) Study 2, (b) Kühl and Szech (2017) laboratory experiment, (c) Andreoni and Bernheim (2009), (d) Grossman (2015) P&O condition, (e) List (2007) baseline, Take $1 and Take $5, and (f) Zhang and Ortmann (2013) baseline, Treatment$1 and Treatment$5 dictator decisions.

I will occasionally return to moral proximity with later examples involving non-cardinal measures. In those cases, the direction of the effect of a factor on salience is obvious and

[7] The effect at longer distances of up to 6000 miles is likely confounded by other factors. Participants report lower feelings of responsibility toward more distant recipients, consistent with moral proximity, but their contributions are not significantly related to distance. The authors attribute this to participants failing truly to have constant beliefs across distances despite the authors attempts to hold all else constant, e.g., through claims about similar per capita GDP. That suspicion seems plausible, since Germany, where the study was conducted, has one of the highest per capita GDPs in the world, and their questionnaire results show a much higher focus on people at longer distances being in need. In addition, Germany has one of the lowest levels of inequality in the world, so even if subjects believe that distant recipients really do enjoy the same average income, levels of inequality elsewhere are surely higher, meaning that, ceteris paribus, the distant poor are likely needier than the local poor.

qualitative variables suffice to explain categorical changes attributed to salience.

4. Moral Uncertainty

Uncertainty is present in virtually all economic decisions, and it can often be managed to some degree. Nevertheless, people sometimes use uncertainty as an excuse to avoid costly actions that are otherwise justified on both economic and moral grounds, such as taking steps to address climate change (Finus and Pintassilgo, 2013). Many studies have demonstrated the relevance of uncertainty to economic decision making, including in economics experiments involving moral preferences, e.g., Bolton, Brandts and Ockenfels (2005), Brock, Lange and Ozbay (2013), Cappelen, Konow, Sørensen and Tungodden (2013a), Rey-Biel, Sheremeta and Uler (2018), Van Koten, Ortmann and Babicky (2013), and Zizzo (2003). In particular, the controlled methods of experiments can help show that the reduction in moral conduct with increased uncertainty is associated with moral preferences themselves and cannot be dismissed as being due solely to other forces, such as risk preferences.

Moral uncertainty refers to the depressing effect on moral salience because of uncertainty in the agent's decision context. Specifically, we consider what I will call the "uncertainty game" in which allocations may be randomly determined by an agent or by default, the latter because either the agent is randomly precluded from choosing allocations or because the agent's choice is not randomly chosen for realization. Assumption 5 specifies the assumed relationship between uncertainty in this game to moral salience.

ASSUMPTION 5: In the uncertainty game, the probability of default constitutes non-moral context with measure n, where $n \in [0,1]$. Moral context has some positive measure, $p > 0$, the value of which depends inversely on the sensitivity of σ to n in the selected context. Baseline moral salience, $\bar{\sigma}$, depends inversely on the unfairness of the default.

Thus, this assumption means that the possibility that allocations will not be based on the agent's choice lowers the moral salience of that choice. Specifically, salience is reduced as the probability increases that the agent's choice will not matter and as the default becomes less fair. In this framework, the moral measure, p, represents a certain implicit moral context and a parameter that effectively calibrates the sensitivity of σ to n. Similarly, the assumption about $\bar{\sigma}$ captures the concept that a less fair (fairer) default lowers (raises) moral salience given that $\sigma = (1 - \bar{\sigma}) \cdot \frac{p}{p+n} + \bar{\sigma}$. Assuming a specific form for the utility function, p and $\bar{\sigma}$ might be estimated

empirically, but the theoretical analysis here does not depend on any particular values for these parameters beyond Assumption 5.

Although numerous economics experiments have investigated uncertainty, I am aware of only a small number with designs suitable to the criteria considered here. As usual, the design must involve non-strategic decisions, and probabilities should be manipulated at two levels at a minimum. Some studies that satisfy these conditions must nevertheless be ruled out because their design activates risk preferences (e.g., Krawczyk and Le Lec, 2010) or fairness preferences over risk because subjects choose levels of risk-taking (e.g., Cappelen et al., 2013). I focus on two studies that satisfy all requirements while representing two different and important categories of moral uncertainty. They lead to the following stylized fact and theorem.

SF/THEOREM 4.1: Dictator transfers decrease at a decreasing rate with the probability of the default. The fairer the default, the greater the transfer (Andreoni and Bernheim, 2009, Grossman, 2015, P&O treatment).

PROOF: The claims about transfers and n follow directly from Assumption 5 and Theorem 2.2 under the assumptions stated there. Writing $x(\sigma(p,n,\bar{\sigma}))$, $\frac{\partial x}{\partial \bar{\sigma}} = \frac{\partial x}{\partial \sigma}\frac{\partial \sigma}{\partial \bar{\sigma}} > 0$, since $\frac{\partial x}{\partial \sigma} > 0$ by Theorem 2.1 and $\frac{\partial \sigma}{\partial \bar{\sigma}} = \frac{n}{p+n} > 0$.

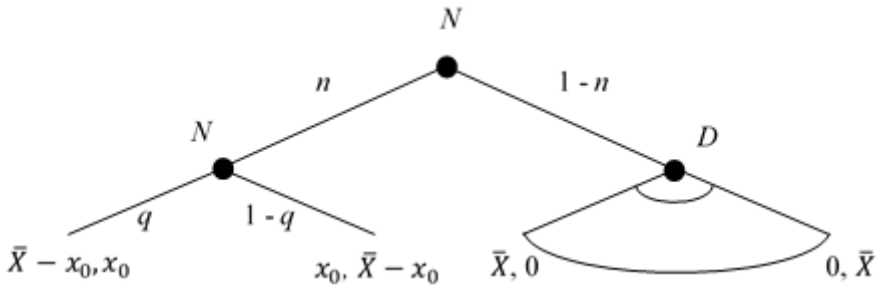


FIGURE 3. – Uncertainty game of Andreoni and Bernheim (2009).

Andreoni and Bernheim (2009) report the results of a dictator game, the design of which is illustrated in extensive form in Figure 3. Nature (N) first decides whether the Dictator (D) or Nature will allocate stakes, $\bar{X}$, of \$20 between D and a recipient (R). Nature allocates with probability n, whereby this probability varies across four levels within subjects, viz., $n \in \{0,0.25,0.5,0.75\}$. If Nature allocates, there is an equal chance, $q = 0.5$, either that $\bar{X} - x_0$ goes to D and x_0 to R or that x_0 goes to D and $\bar{X} - x_0$ to R, where $x_0 \in \{0,1\}$ is varied between subjects. With probability $1 - n$, the allocations will follow the decision of D, who can choose any amount, $x \in X = [0,20]$. Consider panel c of Figure 2, which was introduced in the prior

section and shows the results of logarithmic regressions of the fraction of stakes transferred by D to R on n for $x_0 = 0$ and $x_0 = 1$, separately. The slope and rate of change in fractional transfers are consistent with the predictions: transfers decrease with the probability of the default at a decreasing rate. The pattern of transfers is also consistent with the predicted effect of the fairness of the default: the curve for the fairer $x_0 = 1$ treatment lies above the one for $x_0 = 0$.

Note that, if D finds him/herself in this branch of the game tree, the decision is entirely ex post. That is, the uncertainty has been resolved, and D knows with certainty that his/her transfer will be realized. This type of uncertainty should not matter in standard social preference theories. Andreoni and Bernheim make a persuasive case, however, for their theory of social image that is consistent with the results of this experiment. My aim here is not to fault the social image argument, which I find credible in this instance, indeed, the two accounts are not mutually exclusive. I note now six strengths of an explanation based on moral salience.

First, the moral uncertainty argument does not require that the agent's action set, X, be subject to uncertainty, because moral salience can be affected by information about uncertainty in the decision context, Y. This explains how non-moral elements of Y can reduce moral salience, even when, as in this experiment, decisions are ex post. According to this account, an experiment in which one subject might have been randomly chosen to receive an unfair share of (almost) all of the stakes diminishes the prominence of moral considerations and, therefore, of moral preferences. Thus, it predicts the reduction in giving as the probability of the default increases. Second, and along the same lines, it predicts reduced giving as the default becomes less fair. Moral salience has four additional and attractive features that are unique. Third, it can explain not only the decrease in x with n but also the decreasing rate of change. Fourth, it explains the increase in x with x_0 among those Ds giving more than x_0 due to higher fixed moral salience. This effect is not predicted by, and, in fact, is inconsistent with, social image, which predicts that the higher transfer when $x_0 = 1$ should be due solely to the shift by Ds, who would otherwise give zero at $x_0 = 0$. Fifth, the theory is simple and parsimonious. Sixth, moral salience is consistent with a wide range of other anomalies that are not predicted by alternative accounts such as social image.[8]

Grossman (2015) reports a variation on a dictator experiment designed to test his theory

[8] Moral point salience, which is discussed in Konow (2022a), presents an additional salience-based argument consistent with the higher average transfers when $x_0 = 1$ than when $x_0 = 0$ as well as for masses at those values.

of social-image and self-image. Dictators not only make ex post decisions, which follow the resolution of some uncertainty as in Andreoni and Bernheim, but also face ex ante uncertainty. The experimental design in illustrated is Figure 4. This is a binary dictator game with only two possible pairs of payoffs to D,R of (H,L) or (F,F), where $0 \leq L < F < H$ and $L + H < 2F$, specifically, in Grossman (2015), $H = 7$, $F = 5$, $L = 1$. Nature first determines with equal probability, $q = 0.5$, which of two games the dictator will play, 1 or 2, and this random assignment is common knowledge. These games differ based on whether the default is Fair, i.e., (F,F), in game 1 or Unfair, i.e., (H,L), in game 2. Subjects are randomly assigned one of two probabilities of the default obtaining, $n \in \{0, \frac{2}{3}\}$, which is also common knowledge. Dictators choose the payoffs ex ante in the event their decision is chosen, either the Unfair option A (A1 in game 1 or A2 in game 2) or the Fair option B (B1 in game 1 and B2 in game 2).

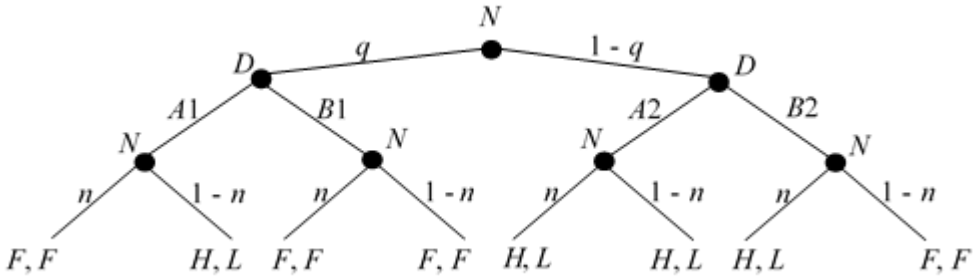


FIGURE 4. – Uncertainty game of Grossman (2015).

Ds make these decisions in three treatments of a between-subjects design that differ with respect to the information that is available to Rs. They can observe the D's choice (C), the outcome (O) or both the probability and outcome (P&O). Comparing variation in the fraction of Fair choices with probabilities in the three treatments, the results produce little support for self-image and mixed results on social image. On the other hand, the results are significantly consistent with most predictions of moral uncertainty in two of three treatments (C, P&O) and insignificantly opposite it in the third (O). This last fact is perhaps because social image concerns muddy the waters somewhat, especially, where they are predicted to do so in O. The results of the treatment with the information conditions closest to the standard dictator game (viz., P&O) are summarized in panel d of Figure 2. This shows that the average transfer to R, or equivalently here, the fraction of Ds choosing Fair, decreases with *n*, and the fairer default results in higher average transfers and a flatter slope. Note that there is one claim of Theorem 2.2 to which the Grossman study cannot speak: since these experiments only vary the probability of the default at

two levels, these results cannot shed light on the rate of change of x with n.

5. The Taking Effect

During civil disturbances and natural disasters, otherwise law-abiding citizens sometimes join in looting (e.g., Green, 2007, Khazan, 6/2/2020, *The Atlantic*, Quarantelli and Dynes, 1968). Scholars have offered many explanations for such behavior, but the results of economics experiments have demonstrated that extrinsic incentives, such as reduced expectations of being punished, cannot, at least solely, explain such abandonment of morals, when opportunities to take from others are offered. Consider an anonymous between-subjects dictator game, in which Rs are also endowed but at a lower level than Ds, and Ds are permitted not only to give in a "Give" treatment but also to take in an otherwise equivalent "Take" treatment. The results show that some Ds take money from Rs in the Take version. Of course, this might be due to Ds, who in the Give version are otherwise constrained to a corner solution at zero, but that does not explain the lower fraction of Ds who choose positive transfers in the Take version versus the Give version (Bardsley, 2008, List, 2007). This *taking effect* means that the addition of taking options results in givers being less generous, indeed, some givers become takers.

The taking effect can be seen as one example of a class of anomalies involving the distinction between helping versus harming others (another example is analyzed in the next section). I make the following general assumption about this class of anomalies.

ASSUMPTION 6: In contexts where $p > 0$, moral salience is increasing the set of helping choices, which involves increasing the payoffs of others, and $n \geq 0$ and is increasing the set of harming choices, which involves decreasing the payoffs of others.

The allocative preference that relates most directly to helping and harming is altruism, so some explanations rest on differences across agents in altruism. Although altruism salience is plausibly more important than fairness salience for explaining helping and harming anomalies, that is not a necessary assumption for the claims except for lesser part of one theorem in the next section.

In dictator games with taking, giving is moral context, and taking is non-moral context. The following assumption fleshes out Assumption 6 for such games.

ASSUMPTION 7: Consider a dictator game where Ds and Rs are endowed with X and Y ($X > Y > 0$), and where giving options may be in any amounts from 0 to the maximum and taking options in any amounts from 0 to the minimum. For concreteness, suppose the moral measure is

$m(C_i) = \max\{c_i \in C_i\} - \min\{c_i \in C_i\}$, $C_i = \{C_+, C_-\}$, 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.

I summarize below many of the rich findings from experiments with taking and propose explanations for them based on moral salience.

SF/THEOREM 5.1: Consider a standard between-subjects dictator game with endowed Ds and Rs, where $X > Y > 0$. Adding taking options to this game reduces giving on both the intensive and extensive margins, i.e., the mean transfer and the frequency of positive transfers fall (e.g., Cappelen et al., 2013b, Cox et al., 2019). Mean transfers fall with taking options at a decreasing rate, i.e., less than proportionately (Bardsley, 2008, Korenok, Millner and Razzolini, 2014, List, 2007, Zhang and Ortmann, 2013). The taking effect diminishes, if the D's choice is observable to the experimenter and other subjects (Alevy, Jeffries, and Lu, 2014).

PROOF: By Assumption 7, adding taking options reduces moral salience, say, from σ^h to σ^l. By Theorem 2.1, all dictators for whom $x > 0$ under σ^h, transfer a lower amount under σ^l. Those, whose preferred transfer is constrained at zero under σ^h, take under σ^l. The effect on transfers is less than proportional with taking options from Theorem 2.2. The effect of observability follows from Assumption 4 and Theorem 2.1.

The results of two studies that vary taking options are presented in panels e and f of Figure 2 (List, 2007, Zhang and Ortmann, 2013, respectively). These illustrate a less than proportionate decline in mean transfers with taking options.

Some dictator experiments vary the endowments of Ds and Rs along with giving and taking options. Specifically, several allow comparisons between a giving game, that is, a standard dictator game where the total endowment, M, is initially all given to the D ($X = M, Y = 0$) and giving is unrestricted, with a taking game, in which M is provisionally allocated to the R ($Y = M, X = 0$) and D taking is unrestricted. Consider now some stylized facts of such games.

SF 5.2: In a between-subjects design, where subjects choose under only one condition, R payoffs ($\pi_R = Y + x$) do not differ significantly between the giving and taking games (Chowdury, Jeon, and Saha, 2017, Dreber et al., 2013, Grossman and Eckel, 2015, Korenok, Millner, and Razzolini, 2014, Smith, 2015). In a within-subjects design, where subjects choose under both conditions, R payoffs are lower in the giving game than the taking game, and, given a choice between playing the giving or taking game, most Ds

prefer the giving game (86% in Korenok, Millner and Razzolini, 2018).

The utility functions of Ds in the giving and taking games can be written, respectively, as

$U^G = u(M - x) + \sigma\phi f(x - \eta) + \sigma g(x, \alpha)$,

$U^T = u(M - x) + \sigma\phi f(x - \eta) + \sigma g(x - M, \alpha)$.

That is, for a given *x*, the D's utility is the same except for the final altruism terms, which reflect the utility gain from giving from the agent's endowment, in the first case, versus the utility loss from taking from the patient's endowment, in the second. This leads to the following theorem.

THEOREM 5.2: In a between-subjects design, it is indeterminate, whether R payoffs will be higher in the giving or the taking game. In a within-subjects design, mean R payoffs are higher in the taking game, and, given the choice between the two games, altruistic Ds, who constitute the majority, prefer the giving game, whereas spiteful Ds prefer the taking game.

PROOF: In a within-subjects design, salience is the same for both decisions by Assumption 1.

By the concavity of g, ${\partial g(x - M, \alpha)}/{\partial x} > {\partial g(x, \alpha)}/{\partial x}$, which implies a larger x, and, therefore, larger R payoffs, in the taking game. By inspection, $U^G|_{\alpha>0} > U^T|_{\alpha>0}$ and $U^G|_{\alpha<0} < U^T|_{\alpha<0}$, meaning altruistic (spiteful) Ds prefer the giving (taking) game. In a between-subjects design, salience is lower in the taking game due to the high non-moral context ($\sigma^T < \sigma^G$), which, ceteris paribus, implies a lower x by Theorem 2.1. Thus, the two effects operate in opposite directions in a between-subjects design such that the overall effect on giving is theoretically indeterminate.

Thus, in the absence of a salience effect, transfers should be larger in the taking game, which is, in fact, what one observes in the within-subjects design. But when moral salience is lower in the between-subjects taking game, the effect on π_R is ambiguous. Although indeterminacy is a nonspecific prediction, it is inconsistent with theory in the absence of salience, and it is consistent with the insignificant differences in π_R in this case compared to the higher π_R in the taking game in the within-subjects design. Note that this theorem implies only spiteful agents prefer the taking game, which the Korenok et al. study cited in SF 5.2 implies is at 14%.

Numerous explanations have been offered for the taking effect. Bardsley (2008) conjectures that it is an experimental artefact, viz., an experimenter demand effect, that is, a desire to please the experimenter. In this context, subjects view the offered choice set as signaling what the experimenter wishes the subject to do. But in a recent and rigorous analysis of

experimenter demand effects, de Quidt, Haushofer and Roth (2018) find such effects to be modest, and they do not seem plausibly to explain the large magnitude of the taking effect. Cappelen et al. (2013b) test whether the choice set signals entitlements, e.g., a taking opportunity might signal the D is morally entitled to do so. But they find that reinforcing entitlements with a real task has no significant effect while the taking effect remains. Korenok, Millner and Razzolini (2012, 2018) point to warm glow and taking aversion (or an endowment effect) and, indeed, altruism in the current model is equivalent to the melding of these two effects. Nevertheless, that alone does not explain the between- versus within-subjects differences. Alevy et al. (2014) argue their results on observability and gender are consistent with social- and self-signaling. As previously discussed, I consider signaling arguments credible, but the present framework is offered as a simpler account, which also explains the observability effect in terms of moral salience, specifically, moral proximity.

List (2007) proposes a "moral cost function," which Cox et al. (2019) formalize. They propose and test experimentally a theory with moral reference points that depend on choice sets. Despite differences in theoretical formulation and some differences in predictions, I view the current project as having points in common with Cox et al., which along with Kimbrough and Vostroknutov (2016) and Krupka and Weber (2013), underscore the importance of the changes in agent sensitivity to the violation of moral norms based on differences in choice sets. Whereas these approaches assume certain changes in norms and sensitivity to norms, the present theory derives these and other patterns from a general theory of stable moral norms and context-dependent moral salience.

6. Joy of Destruction

People sometimes destroy the wealth of others at a personal cost, often risking punishment and obtaining no material benefit to themselves, e.g., some people vandalize property or write computer viruses. There are examples of people, who cooperate over generations but suddenly begin engaging in destructive behavior toward one another. For instance, Serbs, Croats and Bosniaks lived peaceably, often intermarrying, prior to the breakup of Yugoslavia, but subsequently turned on one another, and over 100,000 lives were lost and vast amounts of property destroyed in the Bosnian War. Depending on the particular case, such behavior might be attributed to ethnic hostility, preemptive retaliation, revenge, etc. Economics experiments, however, have documented that, even when such motives can be ruled out by

design, some people are willing to incur a cost to destroy the wealth of others and that such behavior can be easily triggered.

Various "money-burning" games have found that up to almost one-half of subjects acting individually in simultaneous games with unequal endowments destroy earnings of other members of their group, e.g., Zizzo and Oswald (2001), Abbink and Sadrieh (2009), Abbink and Hermann (2011). Similar behavior is observed, when players interact over multiple periods in so-called "vendetta" games, e.g., Abbink and Herrmann (2009), Bolle, Tan and Zizzo (2014). In these studies, however, one cannot rule out motives other than a pure desire to destroy. When endowments are unequal, subjects can be motivated by inequality aversion to destroy. Moreover, as we will see, relatively few subjects destroy in non-strategic decisions (13-15%), but a much higher percentage expect others to destroy (38% in Abbink and Herrmann, 2011), which is consistent with preemptive retaliation in these experiments. Thus, we will focus, as usual, on simple, non-strategic decisions in the cases that follow, such as non-strategic versions of the "joy-of-destruction" (or JD) game, which resembles a dictator game, in that it is unilateral, but with options to destroy others' endowments. In the standard version, endowments are equal, and agents can destroy at zero cost, and zero benefit, to themselves.[9] The utility function of the agent in a non-strategic JD game can be written

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

In the standard JD game, $x \leq 0$, the context is simple and stakes are fixed, so we assume equal endowments, i.e., $X = Y = \eta = M/2$, where $M = X + Y$. We will then consider cases where endowments are unequal and the agent may destroy or create money for the patient, i.e., x can be positive or negative.

SF 6.1: In the standard non-strategic JD game with symmetric endowments, a minority of agents engages in destruction (13% in Iriberri and Rey-Biel, 2013, 15% in Kessler, Ruiz-Martos and Skuse, 2012, and only 13% even in the strategic game of Abbink and Herrmann, 2009).

THEOREM 6.1: In the standard non-strategic JD game with $X = Y$, only a minority, consisting solely of spiteful agents, destroys.

[9] Thus, the agent's endowment is fixed in most JD games, as we assume in the current analysis. Although it shares this feature with spectator decisions, this is, nonetheless, treated as a stakeholder decision and, therefore, the altruism term is included in the agent's utility function. This is because the JD context casts the agent in a personal, agent-patient relationship, similar to a dictator game, and not as a spectator choosing impartially for others.

PROOF: See Appendix.

If the claim of Theorem 6.1 that only spiteful subjects destroy is correct, this implies 13-15% of subjects in the studies cited in SF 6.1 are spiteful. Note that these percentages are not only very consistent with one another across variations in this design but also are remarkably close to the estimate of 14% spiteful agents cited in Section 5 using a different design.

SF 6.2: In JD games, destruction is directed mostly toward advantaged subjects, when inequalities are unfair. That is, it is directed toward richer subjects, when endowments are unearned (Zizzo, 2003, Zhang and Ortmann, 2013), but, when endowments are earned, destruction is mostly aimed at richer subjects only if inequalities are unfair (Fehr, 2018).

THEOREM 6.2: In JD games, suppose fairness is based at least in part on an equity principle, according to which the entitlement is increasing in earned contributions and equal for unearned ones (Konow, 2000). Then, destruction is directed mostly toward those with unfairly high endowments, i.e., richer subjects in games with unearned endowments, and, in games with earned and unequal endowments, mostly toward richer subjects only if inequalities are unfair. Specifically, altruistic and selfish agents, who comprise the majority, only destroy earnings of unfairly advantaged patients, whereas spiteful agents also destroy earnings of some unfairly disadvantaged patients. The richer the unfairly advantaged patient, the more agents of all types destroy.

PROOF: See Appendix.

These findings underscore the importance of inequity aversion, whether the entitlement is equal or not, as opposed to spite alone, in explaining much of the destruction in JD, money burning and vendetta games such that even some altruistic subjects might engage in destruction.

7. Moral Egress

Some people will give money to a beggar but prefer to cross the street, if possible, to avoid the beggar. Field experiments have established that avoidance of this kind is widespread. Forewarning people of door-to-door charitable solicitations results in a large and significant drop in the fraction of homeowners, who open their door at the pre-announced hour, compared to those who are not forewarned and who give more generously and at a higher rate (DellaVigna, List, and Malmendier, 2012). Placing Salvation Army bell-ringers at both of two entrances to a supermarket, rather than just one that can be avoided, increases both the rate and level of donations (Andreoni, Rao, and Trachtman, 2017). Laboratory experiments have found these

results to be robust to controls for possible extrinsic motives, such as social pressure or social image concerns. I call this anomaly *moral egress*: people comply with moral norms, when the norms are salient and exit is prohibitively costly or impossible, but many prefer to exit a situation with high moral salience, when possible.

Moral egress is one example of a class of widely studied anomalies I call "norm avoidance" that involve situations in which agents may choose not only transfers but context itself and, thereby, affect the salience of moral norms. I consider two other examples of norm avoidance to this class of decisions, viz., willful ignorance and delegation, in Konow (2022a). I add the following assumption to examples of norm avoidance (this appears as Assumption 1 in the companion paper).

ASSUMPTION 8: 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 on or acquiring information about the consequences of one's action (σ^m), even if the agent does not exercise that 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, we assume $\sigma^h > \sigma^m > \sigma^l > 0$.

Dana, Cain and Dawes (2006) introduced an experiment that sheds light moral egress. Dictators first play a standard dictator game with \$10. Then they are told for the first time that they may either implement the division or exit the game with \$9, in which case the Rs receive nothing and never find out about the D decision. In the standard game, mean D transfers are at the usual level of about 25% of the endowment, but up to 43% of Ds choose instead to exit and take the \$9. Exit is inconsistent with standard social preference models: selfish Ds should stay in the game and take \$10 instead of \$9, whereas fair-minded Ds should also stay but share fairly. The discussion here centers on a version of this experiment by Lazear, Malmendier and Weber (2012) that extends the Dana et al. design and lends itself to further analysis. This version, which I will call the "exit game," is illustrated in Figure 5. Panel a shows their "no sorting" treatment, which is a standard D game, and panel b illustrates a "sorting" treatment. In the sorting treatment, Ds first choose whether to enter or to exit the D game. They know that, if they enter, they will then proceed to make a decision about how much to transfer of their endowment, X. If they choose to exit, they receive X less a cost c, where $0 \leq c < \eta = \frac{1}{2}X$. That is, in some variations, the exit cost is zero, but, when positive, it is smaller than the common entitlement,

which, as usual in simple D games with fixed stakes, I take to be equal splits.

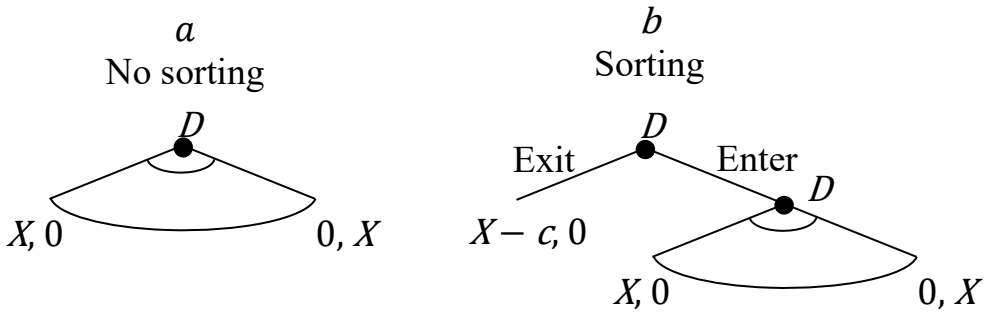


FIGURE 5. – Exit game of Lazear, Malmendier and Weber (2012).

Consider the following stylized facts from some experiments with exit options.

SF 7.1: In dictator games with exit, some dictators enter and transfer zero, some enter and transfer a positive amount, and some exit. Those who exit are, on average, more generous types than those who enter. Mean transfers in the exit game are lower than in the standard dictator game without exit (Broberg, Ellingsen, and Johannesson, 2007, Dana, Cain and Dawes, 2006, Lazear, Malmendier, and Weber, 2012).

For the analysis, Assumption 8 means that, relative to moral salience in the standard dictator game with no exit (σ^h), the availability of an exit option lowers moral salience for those, who elect to enter (σ^m), and moral salience is even lower for those who exit and avoid the dictator decision altogether (σ^l). Denote the utility of a D who enters the game UN, one who exits UX, and one in a standard dictator game without any exit option UD. Then, their utility functions are

$$UD = u(X - x) + \sigma^h \phi f(x - \eta) + \sigma^h g(x, \alpha),$$
$$UN = u(X - x) + \sigma^m \phi f(x - \eta) + \sigma^m g(x, \alpha),$$
$$UX = u(X - c) + \sigma^l \phi f(-\eta).$$

The following theorem explains the relationships between dictator choices in the exit game based on salience and dictators' moral preferences.

THEOREM 7.1: In the exit game, the least fair and least altruistic dictators enter and transfer zero. More altruistic dictators are more likely to enter and transfer a positive amount than to exit or to enter and transfer zero. Fairer dictators are more likely either to exit or to enter and transfer a positive amount than to enter and transfer zero. Specifically, the fairest dictators prefer to exit (enter and transfer zero), if the percentage reduction in salience is greater (less) than the percentage increase in inequity aversion.

PROOF: See Appendix.

Thus, the model indicates that those who enter and transfer zero are the least altruistic and least fair dictators. Higher altruism leads to fewer zero transfers and less exit. Fairer dictators are more likely to exit or to enter and transfer a positive amount, but how ϕ affects which of these is chosen depends: if exiting reduces salience more than it increases inequity aversion (in percentage terms), then the fairest dictators exit, an interpretation that is consistent with SF 7.1 and with generosity, in this game, being driven chiefly by fairness.

The following stylized fact compares mean transfers in the exit game with the standard dictator game.

SF 7.2: In a dictator game with exit, mean transfers are lower than in the standard dictator game without exit (Broberg, Ellingsen, and Johannesson, 2007, Dana et al., 2006, Lazear et al., 2012).

The theory is consistent with this SF for the different experimental designs with exit, but it is worked out formally for the exit game in the following theorem.

THEOREM 7.2: In the exit game, mean transfers are lower among those who enter than in the standard dictator game without exit.

PROOF: Moral salience is lower in the exit game (σ^m) than in the standard game without exit (σ^h), so the optimal transfer x is lower in the former by Theorem 2.1. Moreover, higher exit among the most generous dictators, according to SF 1.1, reinforces the reduced giving in the exit game.

Other findings from games with exit can be explained by changes in moral salience. Lazear et al. (2012) and Andreoni et al. (2017) find that exit increases, if agents must confront patients face-to-face, and Dana et al. (2006) find that both exit and transfers fall, if Rs are never told there was a dictator game regardless of whether the D exits. As described in section 4, such procedural differences are expected to affect moral salience in the form of moral proximity: moral salience increases with personal knowledge of the agent and even of the existence of the agent in a capacity that can affect the patient. Another finding consistent with this model comes from Lazear et al. (2012), who report a result from an exit game, which is stated and proven in the following theorem.

THEOREM 7.3: In an exit game with a constant value of exit ($\bar{X} - c$), the frequency of exit decreases as the stakes of entering the game (X) increase, assuming quite generally that

$0 \leq \frac{d\eta}{dX} \leq 1$.

PROOF: See Appendix.

This result is due to the fact that, as the stakes rise, the utility from entering increases.

There are explanations other than moral salience for some of the experimental results on exit. For instance, if the D chooses to exit, Rs do not find out about the dictator game in these studies, which raises the question of whether Ds are responding to other forces such as social image concerns or guilt aversion, i.e., disutility from giving less than what the R expects. On the latter effect, the evidence is mixed, e.g., Charness and Dufwenberg (2006) find support for guilt aversion in a strategic game with communication, whereas Ellingsen et al. (2010) find the effects are close to zero in three games, including the dictator game. Regarding the former effect, I am unaware of any tests of social image in experiments with exit, so one cannot rule it out. The relative strengths of the moral egress argument are its simplicity and its position in the broader theoretical framework of moral salience. It rests on the intuition that people sometimes distance themselves from situations in which moral norms are salient, because norm compliance reduces utility. In fact, there is supportive evidence of this, assuming utility corresponds to subjective well-being: Ds, who are paired with Rs but given no opportunity to share their endowment with them, are happier, on average, than Ds in a standard dictator game (Konow, 2010).

8. Classic Results

The theory introduced in this paper to explain anomalies is also consistent with classic results on social preferences, as demonstrated in this section.

SF 8.1: There is a mass at null transfers in the standard dictator game without taking options (e.g., 36%, on average, across multiple studies in the survey of Engel, 2011).

THEOREM 8.1: Suppose for the least fair dictators (i.e., those with $\underline{\phi}$) in the standard game with salience σ^* it is the case that ${\partial u(X)}/{\partial x} > \sigma^* \underline{\phi} \, {\partial f(-\eta)}/{\partial x}$. Then there is a mass at null transfers.

PROOF: In this game, transfers are constrained to be non-negative, so a corner solution results at $x = 0$ among that fraction of dictators who are comparatively self-interested, i.e., for whom ${\partial u}/{\partial x}\,(X) \geq \sigma^* \underline{\phi} \, {\partial f(-\eta)}/{\partial x} + \sigma^* \, {\partial g(0,\alpha)}/{\partial x}$. Specifically, sufficient conditions for this are the mass of dictators, who are both the least fair and not altruistic (i.e., given the assumption

$A(0) > 0$).

SF 8.2: Some dictators in the standard game without taking make "super-fair" transfers, i.e., transfers of more than one-half (e.g., 13%, on average, across various studies in Engel, 2011, 6% in the Standard treatment in Konow, 2010). This is a minority of dictators that is smaller than the fraction of those who make null transfers.

THEOREM 8.2: In the standard dictator game, a minority of dictators makes "super-fair" transfers, which are not optimal in the absence of altruism.

PROOF: The assumption that $0 < A(\overline{\alpha}) - A(\alpha^*) < 0.5$, where $\alpha^* = \{\alpha | {\partial u(X-\eta)}/{\partial \pi_a} = \alpha \cdot {\partial g(\eta,\alpha)}/{\partial x}\}$ in this dictator game, implies there is a minority of dictators, whose optimal transfers are super-fair, since for them ${\partial u(X-\eta)}/{\partial \pi_a} < \sigma^* \underline{\phi}\, {\partial f(0)}/{\partial x} + \sigma^* {\partial g(\eta,\alpha)}/{\partial x}$ when $x = \eta$. Such transfers are never optimal in the absence of the altruism term since ${\partial u(X-\eta)}/{\partial \pi_a} > \sigma^* \underline{\phi}\, {\partial f(0)}/{\partial x} = 0$.

SF/THEOREM 8.3: Consistent with SF 8.1 and SF 8.2, assume that null transfers are more numerous than super-fair transfers. Then the mean transfer in the standard dictator game is strictly between zero and one-half of the stakes (e.g., Camerer, 2003, Engel, 2011).

PROOF: See Appendix.

On the basis of this theorem, one can disregard super-fair dictators, whenever the focus is on the mean behavior of dictators in the standard game.

SF/THEOREM 8.4: In the standard dictator game, some dictators transfer amounts that equalize or come close to equalizing allocations (e.g., Camerer, 2003, Engel, 2011).

PROOF: In a simple dictator game, which lacks information about effort, need or other distributive norms, the entitlement reduces to equal splits according to Assumption 3. Combined with Theorem 8.3, transfers closer to equality in these games are consistent with dictators, who have higher values of ϕ, possibly combined with higher values of α.

Note that strict equality does not emerge from fairness preferences alone, since $\phi < \infty$, but it can with the added effect of altruism.

In many dictator games, recipients are endowed. A design I call the "tax experiment" consists of dictator games in which a fixed total endowment ($\overline{M}$) is distributed differently across treatments between dictator (X) and recipient (Y), where $X > Y$ and $\overline{M} = X + Y$.

SF/THEOREM 8.5: Crowding out is partial (or incomplete) in the tax experiment. Incomplete crowding out means that the average dictator transfer, *x*, decreases by less than any increase in the recipient's endowment (e.g., Bolton and Katok, 1998, Korenok, Millner and Razzolini, 2017, Cox, List, Price, Sadiraj, and Samek, 2019).

PROOF: See Appendix for the proof of incomplete crowding out, i.e., $-1 < {}^{dx}/_{dY} < 0$.

Fairness alone predicts complete crowding out, i.e., ${}^{dx}/_{dY} = -1$, so the presence of altruism generates the partial crowding out in this model.

Another piece of corroborating evidence for including altruism can be found in the study of Crumpler and Grossman (2008). In what I will call the "futile dictator" experiment, the experimenter makes a preset charitable donation, and dictators can also contribute to the charity, but then the experimenter's donation to the charity is reduced by the same amount as the dictator's gift, so that the amount received by the charity remains the same. Nevertheless, most dictators (57%) contribute a significant fraction of their endowment (20%, on average). This result is also consistent with conditional altruism, as proven in the following theorem.

THEOREM 8.6: Some dictators contribute a positive amount in the futile dictator experiment.

PROOF: See Appendix.

Such transfers cannot be explained by fairness but are consistent with agents, whose altruism is sufficiently strong. The estimates from Crumpler and Grossman not only provide further support for altruism (or warm glow) but are also consistent with our assumption that the lower bound on the fraction of agents with altruistic preferences ($\alpha > 0$) is greater than one-half.

Recipients are also endowed in what I call the "subsidy experiment," viz., the dictator's endowment is held constant ($\bar{X}$) while the recipient's endowment is varied across treatments.

SF/THEOREM 8.7: Crowding out is partial in the subsidy experiment. Thus, the average dictator transfer, *x*, decreases by less than any increase in the recipient's endowment (e.g., Konow, 2010, Korenok, Millner and Razzolini, 2012).

PROOF: See Appendix for the proof of partial crowding out, i.e., $-1 < {}^{dx}/_{dY} < 0$.

Altruism alone predicts complete crowding out, i.e., ${}^{dx}/_{dY} = 0$, so this validates the inclusion of fairness in the agent's moral preferences.

SF/THEOREM 8.8: When moral norms are activated by the availability of choices (i.e., context *X*) or by information (i.e., context *Y*), spectator and stakeholder allocations are

significantly positively related to affected norms that involve inequality, including equity/proportionality (e.g., Cherry, Frykblom, and Shogren, 2002, Konow 2000, Konow, Saijo and Akai, 2020, Korenok, Millner and Razzolini, 2017, Oxoby and Spraggon, 2008), need (e.g., Benz and Meier, 2008, Eckel and Grossman, 1996, Konow, 2010, 2019, Müller and Renes, 2021, Traub and Kittel, 2020) and efficiency (Almås, Cappelen, and Tungodden, 2020, Charness and Rabin, 2002, Engelmann and Strobel, 2004, Faravelli 2007).

PROOF: This follows from Assumption 3 and Theorem 2.5.

9. Conclusions

This paper proposes a tractable theory that explains both classic results on allocative preferences as well as a wide range of anomalous findings about moral behavior, including moral proximity, moral uncertainty, the taking effect, joy of destruction, and moral egress. 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), moral identity (Bénabou and Tirole, 2011), and social 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 arguably unprecedented set of enigmatic empirical results on moral preferences.

Future 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 to avoid factors that might confound inferences about the forces being studied. But further work might extend the theory to situations involving strategic interaction, such as bargaining.

The moral preferences treated in the current paper are allocative preferences, but evidence from many experiments indicate that agents are also motivated to sanction others, i.e.,

to reward or punish others for their compliance or non-compliance, respectively, with allocative preferences. Konow (2022a) presents new theoretical and empirical findings that build on the current paper. It introduces a new theory of sanctioning, which is related to virtue ethics and is called virtue preferences, and combines it with the theory of moral salience and conditional altruism. This generalized theory is applied to explain classic findings on reward and punishment called reciprocity as well as additional anomalies, including outcome bias (i.e., sanctioning others for their uncontrollable luck), willful ignorance, and delegation. That paper also reports the results of an original experiment that tests the theory out-of-sample and proves consistent with the general theory, including with results discussed in both papers. Finally, whereas this paper employs moral set salience, which involves subsets of context, Konow (2022a) also discusses moral point salience, which involves individual elements of the context, to account for masses at certain decisions. The complete theory extends the range of classic and anomalous findings that can be explained and also provides guidance on distinguishing the contexts in which a social preferences approach suffices and when it is necessary to extend it to take account of moral salience and virtue preferences.

REFERENCES

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.

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.

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

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.

Bénabou, Roland and Jean Tirole (2006): "Belief in a Just World and Redistributive Politics," *Quarterly Journal of Economics,* 121(2), 1652-1678.

Bénabou, Roland and Jean Tirole (2011): "Identity, Morals, and Taboos: Beliefs as Assets," *Quarterly Journal of Economics,* 126(2), 699-746.

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.

Bergh, Andreas, and Philipp C. Wichardt (2018): "Mine, Ours or Yours? Unintended Framing Effects in Dictator Games," *CESifo Working Paper No. 7049*, 1-16.

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.
Bolle, Friedel, Johnathan H.W. Tan, and Daniel John Zizzo (2014): "Vendettas," *American Economic Journal: Microeconomics,* 6(2), 93-130.
Bolton, Gary E., Jordi Brandts, and Axel Ockenfels (2005): "Fair Procedures: Evidence from Games Involving Lotteries," *The Economic Journal,* 115(506), 1054-1076.
Bolton, 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.
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.
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 (2021): "Prosociality Predicts Health Behaviors during the COVID-19 Pandemic," *Journal of Public Economics,* 195, #104367, 1-23.
Cappelen, Alexander W., Astri Drange Hole, Erik O. Sorensen, and Bertil Tungodden (2007): "The Pluralism of Fairness Ideals: An Experimental Approach," *American Economic Review,* 97(3), 818-827.
Cappelen, Alexander W., James Konow, Erik O. Sorensen, and Bertil Tungodden (2013a): "Just Luck: An Experimental Study of Risk-Taking and Fairness," *American Economic Review,* 103(4), 1398-1413.
Cappelen, Alexander W., Ulrik H. Nielsen, Erik O. Sorensen, Bertil Tungodden, and Jean-Robert Tyran (2013b): "Give and Take in Dictator Games," *Economics Letters,* 118(2), 280-283.
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.
Cherry, Todd L., Peter Frykblom, and Jason F. Shogren (2002): "Hardnose the Dictator," *The American Economic Review,* 92(4), 1218-1221.
Chetty, Raj, Adam Looney, and Kory Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review,* 99(4), 1145-1177.
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.

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.
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 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, Ernest and Klaus M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics,* 114(3), 817-868.
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.
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, Michael I. Norton. And Roberto A. Weber (2016): "Motivated Bayesians: Feeling Moral While Acting Egoistically," *Journal of Economic Perspectives,* 30(3), 189-212.
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 (2015): "Self-Signaling and Social-Signaling in Giving," *Journal of Economic Behavior and Organization,* 117, 26-39.
Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982): "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization,* 3(4), 367-388.
Hoffman, Elizabeth, Kevin McCabe, Keith Shachet, and Vernon L. Smith (1994): "Preferences, Property Rights, and Anonymity in Bargaining Games," *Games and Economic Behavior,* 7(3), 346-380.
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.
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.
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 Vostroknutov (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, Tatsuyoshi Saijo, and Kenju Akai (2020): "Equity Versus Equality: Spectators, Stakeholders and Groups," *Journal of Economic Psychology,* 77, 1-28.
Konow, James (2022a): "Virtue Preferences: Jekyll and Hyde Paradoxes with Sanctions," working paper.
Konow, James (2022b): "Jekyll and Hyde (Non-)Paradoxes: The Bystander and Identifiable Victim Effects," work-in-progress.
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.
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.
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 (2021): "Fairness Views and Political Preferences: Evidence from a Large and Heterogeneous Sample," *Social Choice and Welfare*, 56, 679-711.
Oxoby, Robert J. and John Spraggon (2008): "Mine and Yours: Property Rights in Dictator Games," *Journal of Economic Behavior,* 65(3-4), 703-713.
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.
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.
Singer, Peter (2013): "Heartwarming Causes are Nice, but Let's Give to Charity with Our Heads," *The Washington Post,* Opinions, December 19, 2013.
Smith, Alexander (2015): "On the Nature of Pessimism in Taking and Giving Games," *Journal of Behavioral and Experimental Economics,* 54, 50-57.
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.
Traub, Stefan, and Bernhard Kittel, eds. (2020): *Need-based Distributive Justice*, Cham, Switzerland: Springer Press.
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.
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: Proofs

Proof of Theorem 2.1:

$$dU/dx = -u_x(X - x) + \sigma\phi f_x(Y + x - \eta) + \sigma g_x(x, \alpha) = 0$$

letting subscripts on terms of the utility function denote partial derivatives with respect to the subscripted variable(s), e.g., $f_x \equiv \partial f/\partial x$ and $f_{xx} \equiv \partial^2 f/\partial x^2$. Applying the implicit function theorem to solve for $x(\sigma)$, substituting into the first order condition, and differentiating with respect to σ,

$$u_{xx}\frac{dx}{d\sigma} + \phi f_x + \sigma\phi f_{xx}\frac{dx}{d\sigma} + g_x + \sigma g_{xx}\frac{dx}{d\sigma} = 0$$

or, rearranging,

$$dx/d\sigma = \frac{-\phi f_x - g_x}{u_{xx} + \sigma\phi f_{xx} + \sigma g_{xx}} > 0,$$

since from the first order condition above $\phi f_x + g_x = u_x/\sigma > 0$.

Proof of Theorem 2.2

Noting $x(\sigma)$, we can write the composite function $x\big(\sigma(p, n)\big)$. By the chain rule

since $\frac{dx}{d\sigma} > 0$ by Theorem 2.1 and $\frac{\partial\sigma}{\partial n} < 0$ by assumption.

Taking the second derivative,

$$\frac{\partial^2 x}{\partial n^2} = \frac{d^2x}{d\sigma^2}\left(\frac{\partial\sigma}{\partial n}\right)^2 + \frac{dx}{d\sigma}\frac{\partial^2\sigma}{\partial n^2} > 0$$

since, by assumption, $\frac{d^2x}{d\sigma^2} \geq 0$ and $\frac{\partial^2\sigma}{\partial n^2} > 0$ given $p > 0$.

Proof of Theorem 2.3

Solving for $x(\phi)$, substituting, and proceeding as before,

$$u_{xx}\frac{dx}{d\phi} + \sigma f_x + \sigma\phi f_{xx}\frac{dx}{d\phi} + \sigma g_{xx}\frac{dx}{d\phi} = 0.$$

$$dx/d\phi = \frac{-\sigma f_x}{u_{xx} + \sigma\phi f_{xx} + \sigma g_{xx}} \gtreqless 0 \text{ if } f' \gtreqless 0 \text{ as } x \lesseqgtr \eta - Y.$$

Proof of Theorem 2.4

Solving for $x(\alpha)$, substituting, and differentiating,

$$u_{xx}\frac{dx}{d\alpha} + \sigma\phi f_{xx}\frac{dx}{d\alpha} + \sigma g_{xx}\frac{dx}{d\alpha} + \sigma g_{x\alpha} = 0.$$

$$dx/d\alpha = \frac{-\sigma g_{x\alpha}}{u_{xx} + \sigma\phi f_{xx} + \sigma g_{xx}} > 0.$$

Proof of Theorem 2.5

Solving for $x(\eta)$, substituting, and differentiating,

$$u_{xx}\frac{dx}{d\eta} + \sigma\phi f_{xx}\frac{dx}{d\eta} - \sigma\phi f_{xx} + \sigma g_{xx}\frac{dx}{d\eta} = 0.$$

$$0 < \frac{dx}{d\eta} = \frac{\sigma\phi f_{xx}}{u_{xx} + \sigma\phi f_{xx} + \sigma g_{xx}} < 1.$$

Proof of Theorem 6.1

Let $X + Y = M$ and suppose $\eta = \frac{M}{2} = \frac{Y}{2}$. Then

$$U = u(X) + \sigma\phi f(x) + \sigma g(x, \alpha).$$

$$dU/dx = \sigma\phi f_x(x) + \sigma g_x(x, \alpha) = 0$$

$$\Rightarrow \phi f_x(x) = -g_x(x, \alpha)$$

$$\Rightarrow \alpha \gtreqless 0 \Rightarrow g_x \gtreqless 0 \Rightarrow f_x \lesseqgtr 0 \Rightarrow x \gtreqless 0.$$

If x is constrained, $x \leq 0$, then $x = 0$ for $\alpha \geqq 0$ and $x < 0$ only for $\alpha < 0$, which is a minority according to the assumption that $0 < \mathrm{A}(0) < 0.5$.

Proof of Theorem 6.2

The general form for the utility function of the agent in the JD game is

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

The first order condition is

$$dU/dx = \sigma\phi f_x(Y + x - \eta) + \sigma g_x(x, \alpha) = 0$$

or

$$\phi f_x(Y + x - \eta) = -g_x(x, \alpha).$$

It follows that $\alpha \geq 0 \Rightarrow x < 0$ only if $Y > \eta$. When $\alpha < 0$, $x < 0$ if $Y \geq \eta$ as well as for some $Y < \eta$.

Solving for $x(Y)$, substituting it into the first order condition, and differentiating,

$$\sigma\phi f_{xx} + \sigma\phi f_{xx}\frac{dx}{dY} + \sigma g_{xx}\frac{dx}{dY} = 0$$

$$\Rightarrow -1 < dx/dY = -\frac{\phi f_{xx}}{\phi f_{xx} + g_{xx}} < 0$$

for interior solutions, which implies destruction increases with the size of a patient's unfair advantage.

Proof of Theorem 7.1

First, consider those who enter. From Theorem 8.1 that follows, there is a mass of low altruism and low fairness types who transfer nothing, indeed the necessary condition is satisfied a fortiori, since salience is lower than in the standard dictator game, $\sigma^m < \sigma^h$, and, by Theorem 2.1, the optimal transfer here is even lower for all dictators. For those who enter, Theorems 2.3 and 2.4 apply, meaning the optimal transfer, and, therefore, the likelihood of giving a positive amount rather than zero, is increasing in α and in ϕ (note that, although $\partial x/\partial\phi < 0$ for super-fair dictators, this never prompts them to give less than η, let alone zero).

Second, the choice of exiting over entering and transferring zero implies a higher ϕ, since

$$UX = u(X - c) + \sigma^l\phi f(-\eta) > UN = u(X) + \sigma^m\phi f(-\eta),$$

which implies

$$\phi > \phi^X \equiv \frac{u(X) - u(X - c)}{(\sigma^l - \sigma^m)f(-\eta)} > 0.$$

That is, in exit games with a positive exit cost ($c > 0$), there is a strictly positive fairness threshold for exiting such those who exit have stronger fairness preferences than those who enter and transfer zero.

Finally, consider the effects of moral preferences on the choices of exiting versus entering and possibly transferring a positive amount. The effect of α on the utility of exiting is zero since

$$\partial UX/\partial\alpha = 0$$

The effect of α on the utility of entering and transferring a positive amount is

$$\partial UN/\partial\alpha = -u_x\frac{\partial x}{\partial\alpha} + \sigma^m\phi f_x\frac{\partial x}{\partial\alpha} + \sigma^m g_x\frac{\partial x}{\partial\alpha} + \sigma^m g_\alpha$$
$$= (-u_x + \sigma^m\phi f_x + \sigma^m g_x)\frac{\partial x}{\partial\alpha} + \sigma^m g_\alpha = \sigma^m g_\alpha$$
$$= \sigma^m g_\alpha > 0$$

since $-u_x + \sigma^m\phi f_x + \sigma^m g_x = 0$ for interior solutions and $\partial x/\partial\alpha = 0$ for corner solutions. Thus, the utility of entering and transferring a positive amount is increasing in α, so the share of dictators doing so is increasing in α, since the utility of exiting is unaffected by α. The effect of ϕ on the utility of exiting is

$$\partial UX/\partial\phi = \sigma^l f(-\eta) < 0.$$

The effect of ϕ on the utility of entering is

$$\partial UN/\partial\phi = (-u_x + \sigma^m\phi f_x + \sigma^m g_x)\frac{\partial x}{\partial\phi} + \sigma^m f(x-\eta) < 0$$
$$= \sigma^m f(x-\eta) < 0$$

When the optimal $x > 0$, exiting becomes more (less) attractive if

$$\partial UX/\partial\phi > (<)\ \partial UN/\partial\phi, \text{ or}$$
$$\Leftrightarrow \sigma^l f(-\eta) > (<)\sigma^m f(x-\eta)$$
$$\Leftrightarrow \frac{\sigma^m - \sigma^l}{\sigma^m} > (<)\frac{f(-\eta) - f(x-\eta)}{f(-\eta)}.$$

Proof of Theorem 7.3

UX is fixed at $u(\bar{X} - c) + \sigma^l\phi f(-\eta)$, since $\bar{X}$ and c are fixed. The first order condition for an interior solution with a dictator who enters is

$$\partial UN/\partial x = -u_x(X - x) + \sigma^m\phi f_x(x-\eta) + \sigma^m g_x(x,\alpha) = 0.$$

Differentiating with respect to $X, x(X)$ and $\eta(X)$,

$$-u_{xx} + u_{xx}\frac{dx}{dX} + \sigma^m\phi f_{xx}\frac{dx}{dX} - \sigma^m\phi f_{xx}\frac{d\eta}{dX} + \sigma^m g_{xx}\frac{dx}{dX} = 0$$
$$\Rightarrow 0 < \frac{dx}{dX} = \frac{u_{xx} + \sigma^m\phi f_{xx}\frac{d\eta}{dX}}{u_{xx} + \sigma^m\phi f_{xx} + \sigma^m g_{xx}} < 1.$$

The effect of X on UN evaluated at the optimal x is

$$\left.\frac{\partial UN}{\partial X}\right|_x = u_x - u_x\frac{dx}{dX} + \sigma^m\phi f_x\frac{dx}{dX} - \sigma^m\phi f_x\frac{d\eta}{dX} + \sigma^m g_x\frac{dx}{dX}$$

$$= u_x - \sigma^m \phi f_x \frac{d\eta}{dX} + (-u_x + \sigma^m \phi f_x + \sigma^m g_x) \frac{dx}{dX}$$
$$= u_x - \sigma^m \phi f_x \frac{d\eta}{dX} > 0$$

since $0 \leq {}^{d\eta}/_{dX} < 1$ and, from the first order condition, $(-u_x + \sigma^m \phi f_x + \sigma^m g_x) = 0$ such that $u_x > \sigma^m \phi f_x$. Thus, as X rises, so does UN, whereas UX stays the same and more Ds choose entry over exit.

Proof of Theorem 8.3:

In the standard dictator game, the minimum and maximum transfers are zero and *X*, respectively. Denote the null transfer x_N, where $x_N = 0$, and its frequency q_N. Denote the average super-fair transfer x_H, where $\frac{1}{2}X < x_H \leq X$, and the frequency of super-fair transfers q_H. Denote the average transfer between 0 and one-half x_G, where $0 < x_G \leq \frac{1}{2}X$, and its frequency q_G. Suppose $q_N, q_G, q_H \in (0,1)$. Finally, note that $q_N + q_G + q_H = 1$, and, according to SF 3.2, $q_H < q_N$. Then the average transfer equals

$$E(x) = q_N \cdot x_N + q_G \cdot x_G + q_H \cdot x_H = q_G \cdot x_G + q_H \cdot x_H.$$

First, note that $E(x) > 0$, since $x_N = 0$ and $q_G > 0, x_G > 0, q_H > 0$, and $x_H > 0$. Next, show $E(x) < \frac{1}{2}X$. Consider x_G at its maximum value $\frac{1}{2}X$, and x_H at its maximum value X. Note that $q_N \cdot x_N + q_H \cdot X = (q_N + q_H)\frac{q_H}{q_N + q_H} \cdot X < (1 - q_G)\frac{1}{2}X$, since $x_N = 0$, $q_N + q_H = 1 - q_G$, and $\frac{q_H}{q_N + q_H} < \frac{1}{2}$ from the fact that $q_H < q_N$. Then, the least upper bound of $E(x)$ is $\frac{1}{2}X$:

$$E(x) = q_G \cdot \frac{1}{2}X + [q_N \cdot x_N + q_H \cdot X] < q_G \cdot \frac{1}{2}X + (1 - q_G)\frac{1}{2}X = \frac{1}{2}X.$$

Proof of Theorem 8.5

$$U = u\left(\overline{M} - Y - x\right) + \sigma\phi f(Y + x - \eta) + \sigma g(x, \alpha).$$
$${}^{dU}/_{dx} = -u_x\left(\overline{M} - Y - x\right) + \sigma\phi f_x(Y + x - \eta) + \sigma g_x(x, \alpha) = 0.$$

Substituting $x(Y)$ and differentiating,

$$u_{xx} + u_{xx}\frac{dx}{dY} + \sigma\phi f_{xx} + \sigma\phi f_{xx}\frac{dx}{dY} + \sigma g_{xx}\frac{dx}{dY} = 0,$$
$$-1 < \frac{dx}{dY} = \frac{-u_{xx} - \sigma\phi f_{xx}}{u_{xx} + \sigma\phi f_{xx} + \sigma g_{xx}} < 0.$$

Note that, in the absence of altruism, $\frac{dx}{dY} = -1$.

Proof of Theorem 8.6

Let *Y* denote the preset experimenter donation and *y* the amount by which the experimenter reduces the recipient's (R's) earnings. Then R earns $Y + x - y = Y$ since $x = y$.

$$U = u(X - x) + \sigma\phi f(Y - \eta) + \sigma g(x, \alpha).$$
$${}^{dU}/_{dx} = -u_x(X - x) + \sigma g_x(x, \alpha) = 0$$

in the case of interior solutions. The assumptions that $A(\overline{\alpha} - \alpha^*) > 0$ where $\alpha^* \equiv \{\alpha | u_x(X - \eta) = \sigma g_x(\eta, \alpha)\} \Longrightarrow u_x(X) < \sigma g_x(0, \alpha) \ni \alpha^* > 0 \forall \alpha > \alpha^*$, who form the fraction

$0.5 > \mathrm{A}(\overline{\alpha} - \alpha^*) > 0$ plus, on the margin of α^*, some α for whom $\alpha^* \geq \alpha > 0$.

Proof of Theorem 8.7

Since the total stakes, $X + Y$, vary, the entitlement, η, would be impacted according to most distributive principles. Given the simple context, equality is a reasonable assumption, but I make the weaker assumption that $\eta = (1-t)\overline{X} + tY$, $0 < t < 1$.

Then $Y + x - \eta = (1-t)\left(Y - \overline{X}\right) + x$, and

$$U = u\left(\overline{X} - x\right) + \sigma\phi f\left((1-t)\left(Y - \overline{X}\right) + x\right) + \sigma g(x, \alpha),$$

$$dU/_{dx} = -u_x\left(\overline{X} - x\right) + \sigma\phi f_x\left((1-t)\left(Y - \overline{X}\right) + x\right) + \sigma g_x(x, \alpha) = 0.$$

Substituting $x(Y)$ and differentiating,

$$u_{xx}\frac{dx}{dY} + \sigma\phi(1-t)f_{xx} + \sigma\phi f_{xx}\frac{dx}{dY} + \sigma g_{xx}\frac{dx}{dY} = 0,$$

$$-1 < \frac{dx}{dY} = \frac{-\sigma\phi f_{xx}(1-t)}{u_{xx} + \sigma\phi f_{xx} + \sigma g_{xx}} < 0.$$

Note that, in the absence of fairness, $\frac{dx}{dY} = 0$.