SOCIAL PREFERENCES OR SACRED VALUES? THEORY AND EVIDENCE OF DEONTOLOGICAL MOTIVATIONS

Daniel L. Chen and Martin Schonger*

Abstract Recent advances in economic theory, largely motivated by experimental findings, have led to the adoption of models of human behavior where a decision-maker not only takes into consideration her own payoff but also others' payoffs and any potential consequences of these payoffs. Investigations of deontological motivations, where a decision-maker makes her choice not only based on the consequences of a decision but also the decision per se have been rare. We propose an experimental method that can detect individual's deontological motivations by varying the probability of the decision-maker's decision having consequences. It uses two states of the world, one where the decision has consequences and one where it has none. A purely consequentialist decision-maker whose preferences satisfy first-order stochastic dominance will choose the decision that leads to the best consequences regardless of the probability of the consequential state. A purely deontological decision-maker is also invariant to the probability. However, a mixed consequentialist-deontological decision-maker's choice changes with the probability. The direction of change indicates how deontological motivations are incorporated into preferences. One implication is that the random lottery incentive method of eliciting preferences may be misleading when moral decisions are involved.

JEL Codes: D6, K2

Keywords: Consequentialism, Deontological Motivations, Normative Commitments, Social Preferences, Revealed Preference, Decision Theory, First Order Stochastic Dominance, Random Lottery Incentive Method

*Daniel L. Chen, Chair of Law and Economics, ETH Zurich, chendan@ethz.ch; Martin Schonger, Law and Economics, ETH Zurich, mschonger@ethz.ch. Latest version available at: http://nber.org/~dlchen/papers/Deontological.pdf. We thank research assistants and numerous colleagues with helpful comments at economics departments at Tel Aviv, Haifa, Bonn, Basel, Norwegian School of Economics, and Oslo; at the Econometric Society, Society for Advancement of Economic Theory, North America Economic Science Association, International Economic Science Association, Society of Labor Economists, European Association for Decision Making, Association for the Study of Religion, Economics, and Culture, Social Choice and Welfare, American Law and Economics Association, Midwest Political Science Association, and Zurich Design Workshop; at law departments at Hebrew University, Lausanne, ETH Zurich, Bar Ilan; and at the philosophy department at Oslo.This project was conducted while Chen received funding from the European Research Council, Ewing Marion Kauffman Foundation, Institute for Humane Studies, John M. Olin Foundation, Swiss National Science Foundation, and the Templeton Foundation.

1. INTRODUCTION

Economics has traditionally taken a consequentialist view of both individual and social behavior, at least if one defines consequentialism broadly, as we do. In the conventional homo oeconomicus model, a decision-maker (DM) only cares about the consequences of her decision in terms of her own payoffs. Behavioral economists have extended this model to allow for a DM to take into account others' payoffs (Rabin 1993; Fehr and Schmidt 1999). Even more recently, models and experimental evidence have shown that people may also care about others' feelings and attitudes that will result from one's decision (McCabe et al. 2003; Falk and Fischbacher 2006; Bénabou and Tirole 2006; Andreoni and Bernheim 2009). Under a broad conception of consequences such as ours, even these non-monetary outcomes are consequences; but note that some of the literature (Gneezy 2005) has a narrower view of consequences and, therefore, use the term "non-consequential" to refer to what we would view as consequences, such as others' feelings. The question of this paper is whether and in which situations the range of motivations considered should be extended further, to allow for non-consequentialist, specifically deontological (internal duty-oriented) motivations. Adam Smith's impartial spectator in The Theory of Moral Sentiments may have been duty-oriented (Smith 1761).[1] Deontological motivations are present when people care about their decisions per se. Any direct or indirect consequence arising from payoffs or others' observation of or inference about the DM's behavior, intentions, or type, we still consider to be consequentialist. In this delineation, we try to adapt major concepts of moral philosophy to economics, and bring the precision of economic methodology, in particular revealed preference, to moral philosophy. Philosophers and legal theorists commonly assume that people have deontological motivations (Greene et al. 2001; Mikhail 2007). Some debate whether people should have deontological motivations (Nagel 1970; Kant 1959) and whether policy-making should take into account deontological motivations (Kaplow and Shavell 2009). We investigate the positive question of whether people have deontological motivations and what are people's deontological motivations.

Deontological motivations do not coincide with what is sometimes called duty in the political science or behavioral economics literature. For example, people may participate in elections even when their vote is not pivotal due to a "duty" to vote (Riker and Ordeshook, 1968).[2] But participation in elections is observable to family members and neighbors, and thus the word duty may only refer to fulfilling one's social obligations, i.e., to be motivated out of concern for the social consequences of one's actions (for evidence that voting may be motivated by social image concerns, see DellaVigna et al. (2013)). Moreover, people might also worry about the extent of the mandate or send a message by voting for a particular candidate. Such expressive voting falls under a broad conception of consequences and is thus not deontological.

[1] "The patriot who lays down his life for... this society, appears to act with the most exact propriety. He appears to view himself in the light in which the impartial spectator naturally and necessarily views him, ... bound at all times to sacrifice and devote himself to the safety, to the service, and even to the glory of the greater ... . But though this sacrifice appears to be perfectly just and proper, we know how difficult it is... and how few people are capable of making it." (Smith 1761).

[2] For recent experimental evidence on expressive voting, see e.g. Feddersen et al. (2009); Shayo and Harel (2012).

A reason for the lack of attention to deontological motivations is likely due to the difficulty of designing studies that can detect and distinguish the presence of non-consequentialist motivations. This paper makes three contributions: It formalizes the notion of consequentialist as well as deontological motivations as properties of preference relations; suggests a method to use revealed preference to detect deontological motivations; and, using that method, provides experimental evidence for the existence and content of deontological motivations, and thereby hopes to contribute to the understanding of human behavior, specifically behavior that might be viewed as morally motivated or constrained.

Before previewing our formal interpretation of the philosophical concepts of consequentialist and deontological moral philosophy, let us review their definitions. Sinnott-Armstrong (2012) define consequentialism as, "the view that normative properties depend only on consequences" and explains that "[c]onsequentialists hold that choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about." Utilitarianism is one example of a consequentialist moral philosophy (Bentham 1791); in fact any welfarist view is consequentialist (Arrow 2012). By contrast, deontological ethics holds that "some choices cannot be justified by their effects—that no matter how morally good their consequences, some choices are morally forbidden." (Alexander and Moore 2012). Immanuel Kant, one of the most famous proponents of deontological ethics, claims that lying is absolutely prohibited, even if the lie brought about great good. Famously, Kant thinks that it is morally impermissible even to lie to an ax-murderer about the whereabouts of a friend whom the former is pursuing (Kant 1797). Virtues ethics, which originates in the work of Plato and Aristotle, would also be among the non-consequentialist motivations we seek to uncover.[3]

We propose the following formalization of these moral philosophies: Consider a decision d, that may cause (a vector of) consequences x. Assume that preference can be represented by a utility function. Then the preferences of a (pure) consequentialist can be represented by a utility function of the form $u = u(x)$. If in addition to consequentialist motivations the DM also has deontological ones, then the utility function takes the form $u = u(x, d)$.[4] At the other extreme, a DM who is purely deontological would have a utility function of the form $u = u(d)$. Or more precisely, since deontological ethics, unlike consequentialism, is supererogatory and thus may not lead to a unique morally permissible decision, they are lexicographic with the first component being based only on d, and the second component being based only on x. But as a shortcut to grasp the intuition, it is helpful to think of a purely deontological DM to have preferences represented by $u = u(d)$.[5] It may seem odd to model deontological motivations by utility functions since one may view "utility" as a consequence, but since ours is a revealed preference approach, we follow the usual economics approach (Friedman and Savage, 1948) of modeling decision-makers' behavior as if they maximized

[3]To economize on terminology, we will only refer to deontological ethics. We also make no distinction between positive and negative duties.

[4]We use the term "motivation" to make clear that preferences can have both components.

[5]It is possible to model purely deontological people as having a different choice set (Nozick 1974). But traditionally a choice set are the objective, external constraints facing a person. We call the internal constraints preferences. Thus, we model deontological moral constraints on the decision-maker as internal constraints, that is, as the first part of preferences in a lexicographic framework.

that objective function and refrain from interpreting the function as standing for utility or happiness.

Since decisions and consequences are closely tied together, it is not obvious how to cleanly distinguish between these motivations. Moral vignettes have been used by philosophers and psychologists to identify deontological motivations. We provide a revealed preference method to identify deontologicalism. A key aspect of our thought experiment is to vary the probability that one's moral decision is consequential (i.e., implemented). For a consequentialist, the optimal decision is independent of the probability that the action will be enacted, because, roughly speaking, the marginal cost (e.g., lost money or time) and marginal benefit (e.g., recipient's well-being) are both equally affected by the probability. For a deontologist, the optimal decision is also independent of the probability, since the duty to make a decision is unaffected by the probability. Only mixtures of both consequentialist and deontological motivations predict changes in behavior as the probability changes.

Our thought experiment can be viewed as testing the joint hypothesis of consequentialism and first-order stochastic dominance (FOSD): If your decision has only a consequence with a certain probability and something outside of your control happens, the probability cannot affect your choice of best action. Behavioral changes due to the reduction in probability of an act being executed would suggest that people care about their decisions even when they are inconsequential. For example, as the likelihood of actually having to pay for the moral decision decreases, the marginal cost to making the moral decision decreases while the feeling of duty (internal non-consequentialist price of the moral decision) to make the moral choice remains the same, so the decision should typically become more moral. Only those who have both motivations and trade them off change their behavior.

The direction of change indicates how consequentialist and deontological motivations are incorporated into preferences. Under additive utility and preferences that are globally convex, a decrease in the probability of being consequential leads to a decrease in marginal cost (external consequentialist price of the moral decision), so to equate marginal costs and marginal benefits, the decision becomes more moral. The decision can become less moral if consequentialist and deontological motivations are non-additive or if deontological motivations are such that utility, from the decision per se, decreases if the DM decides to do more than what duty calls for (for example, if utility functions are not globally convex). Under functional form assumptions, the magnitude of change indicates the location of one's greatest duty and the weights individuals put on consequentialist and deontological motivations.

We operationalize our thought experiment by asking subjects to make a decision to donate to Doctors Without Borders and place their decision in an envelope that is shredded when the decision is not implemented. Thus, no one ever knows the decision in the non-consequential state, not even the experimenter, in case the DM cares about what the experimenter thinks of the decision or how the experimenter will use the decision. The benefit of the decision per se remains even in the non-consequential state. In a second experiment, we correlate revealed preference with vignette decisions: We pair the donation decision with two modified moral trolley vignettes that ask subjects if they would be willing to kill in order to save many. The two vignettes vary in the number of people who would be saved. A decision-maker who changes decisions may be mixed consequentialist

and deontological. We also correlate revealed preference with demographics that may be indicative of the moral tribe to which an individual belongs (Greene 2014). The third experiment collects data to estimate structurally the trade-off between consequentialist and deontological motivations and the location of sacred values. We collect data from a large number of subjects, for very low implementation probabilities, and in an online setting. We find that when the probability of being consequential decreases, individuals become more moral in their decision-making.

Two aspects of the experimental design address possible confounds, where individuals are consequentialist but violate FOSD. First, the DM may wish to target the ex ante utility of individuals (Machina 1989). Such a DM, would donate more as the probability of her decision being implemented declines. We design a treatment arm where the non-consequential state of the world involves the entire sum being donated. Our results are not consistent with targeting of ex ante utility.[6] Second, the decision may be cognitively costly, and so the cost of the decision per se remains even in the non-consequential state. If cognitive costs change with the probability, subjects may be more likely to report their heuristic decision, which may be a more moral one (Rand et al. 2012). Our response to this is twofold. First, we measure time spent making the decision. Second, we formalize a cognition cost explanation and examine its predictions.

Our paper complements Alger and Weibull (2012)'s formal investigation that deontological preferences will be selected for when preferences rather than strategies are the unit of selection; they find that preferences that are a convex combination of homo oeconomicus and homo kantiensis, which is similar to our definition of purely deontological preferences, will be evolutionarily stable. The remainder of the paper is organized as follows. Using a thought experiment, section 2 defines consequentialism, deontologicalism, and mixed motivations as properties of a preference relation and proves that under first-order stochastic dominance, behavior is invariant to the probability. Section 3 describes how we implemented the thought experiment, and section 4 reports the reduced form results and structural estimates of deontological motivations. Section 5 discusses implications for experimental methods and Section 6 concludes.

2. FORMAL INVESTIGATION

2.1. *Thought Experiment*

The idea to identify non-consequentialist motivations by varying the probability of the DM's decision being consequential guides this paper. The DM has a real-valued choice variable d which influences both her own monetary payoff x_1 as well as the payoff x_2 of a recipient R. There are two states of the world, state C and state N. In state C, the DM's decision d fully determines both x_1 and x_2. In state N, both x_1 and x_2 take exogenously given values, and the decision d has no impact at all. Thus, in state C, the decision is consequential, while in state N, it is not. After DM chooses d, nature randomly decides which state is realized. State C occurs with probability $\pi > 0$, state N with probability $1-\pi$. The structure of the game is public, but the decision d is only known to DM. In state N, therefore, R has no way of knowing d, but, in state C, R knows d, indeed he can infer it from x_2. Superscripts indicate the realized state, so that the payoffs are (x_1^C, x_2^C) in state C, and

[6] Our results are also inconsistent with targeting of ex ante giving.

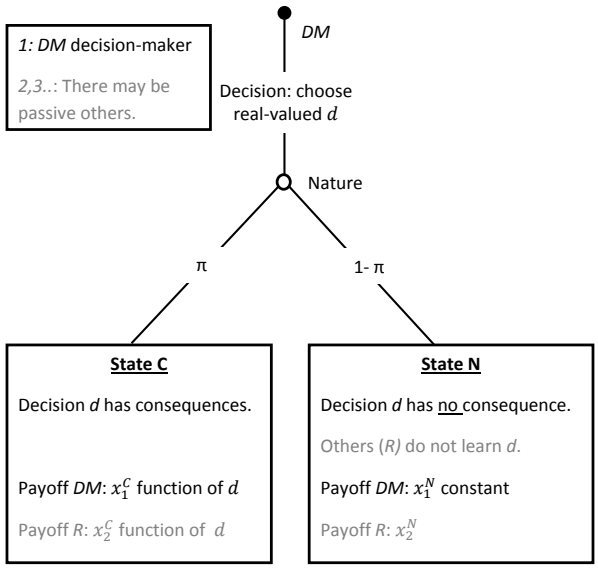


FIGURE 1.—

(x_1^N, x_2^N) in state N. Figure 1 illustrates this.

This general experimental design could be used for many morally relevant decisions; here we apply our identification method to the dictator game and thus to the moral decision to share. As shown in Figure 2, the DM receives an endowment of ω, and must decide how much to give to R. She may choose any d such that $0 \leqslant d \leqslant \omega$ and the resulting payoffs are $x_1^C = \omega - d$ and $x_2^C = d$. For $\pi = 1$, the game thus reduces to the standard dictator game. In state N, a pre-determined, exogenous κ will be implemented, where $0 \leqslant \kappa \leqslant \omega$, and $x_1^N = \omega - \kappa$ and $x_2^N = \kappa$ are the resulting payoffs.

2.2. *A testable implication of the standard framework*

In the following we sketch the standard, consequentialist approach to choice under uncertainty where the central assumption for choice behavior regarding uncertainty is first-order stochastic dominance (FOSD). A wide variety of models of choice under uncertainty satisfies FOSD and thus falls within this framework, among them most prominently, expected utility theory, its generalization by Machina (1982), but also cumulative prospect theory (Tversky and Kahneman, 1992) or rank-dependent utility theory (Quiggin, 1982).

In the following paragraph and the axioms up to FOSD, we closely follow the canonical framework as laid out in Kreps (1988). Let there be outcomes x. x can be a real valued vector. In the thought experiment, it would be $x = (x_1, x_2)$. Let the set of all x be finite and denote it by X. A probability measure on X is a function $p : X \to [0, 1]$ such that $\sum_{x \in X} p(x) = 1$. Let P be the set of all probability measures on X, and therefore, in the thought experiment, a subset of it, is the choice set of the decision-maker. Axiom 1 is the standard one saying that the preference relation is a complete ordering. It implicitly includes consequentialism since the preference relation is on P, that is, over lotteries that are over consequences x.

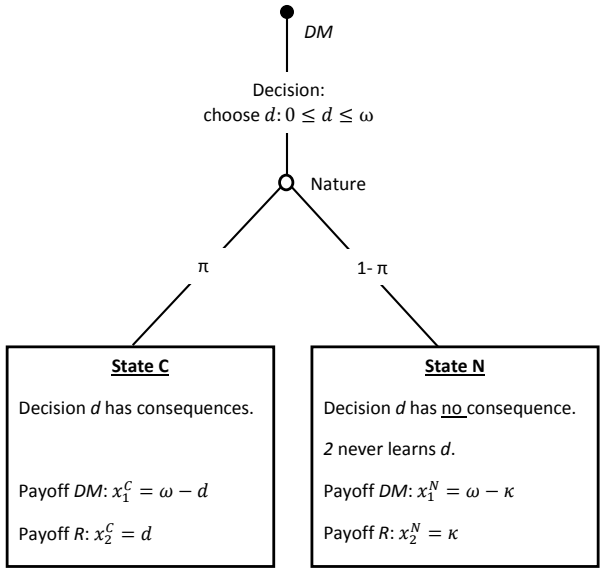


FIGURE 2.—

AXIOM 1 *(preference relation) Let $\succsim$ be a complete and transitive preference on P.*

Next we define first-order stochastic dominance (FOSD). Often, definitions of FOSD are suitable only for preference relations that are monotonic in the real numbers, for example see Levhari et al. (1975). These definitions define FOSD with respect to the ordering induced by the real numbers, assuming that prices are vectors. Such an approach is inappropriate in the context of social preferences, which are often not monotonic due to envy or fairness concerns. For example, Fehr and Schmidt (1999) preferences, which ordinally rank allocations of certain prospects, would violate such definitions of FOSD since they do not satisfy monotonicity and do not convey the DM's attitude about risk.

DEFINITION (FOSD) p first-order stochastically dominates q with respect to the ordering induced by $\succsim$, if for all x':

$$\sum_{x:x' \succsim x} p(x) \leq \sum_{x:x' \succsim x} q(x).$$

AXIOM *(FOSD) If p FOSD q with respect to the ordering induced by $\succsim$, then $p \succsim q$.*

DEFINITION (Strict FOSD) p strictly first-order stochastically dominates q with respect to the ordering induced by $\succsim$ if p FOSD q with respect to that ordering, and there exists an x' such that:

$$\sum_{x:x' \succsim x} p(x) < \sum_{x:x' \succsim x} q(x).$$

AXIOM *(Strict FOSD) If p strictly FOSD q with respect to the ordering induced by $\succsim$, then $p \succ q$.*

The following theorem implies that in our thought experiment changing the probability of being consequential π does not change the decision. It is this prediction of the theory that we will test and interpret a rejection of the prediction as evidence that people are not purely consequentialist.

THEOREM 1 *If the DM satisfies the axioms Preference Relation, FOSD, and Strict FOSD, and there exist* $x, x', x'' \in X'$ *and* $\pi\epsilon(0;1]$ *such that* $\pi x+(1-\pi)x'' \succcurlyeq \pi x'+(1-\pi)x''$*, then for all* $\pi'\epsilon(0;1]$ *:* $\pi'x + (1-\pi')x'' \succcurlyeq \pi'x' + (1-\pi')x''$.

PROOF: (i) $x \succsim x'$: Suppose not, then $x' \succ x$, and therefore $\pi x' + (1-\pi)x''$ strongly first-order stochastically dominates $\pi x+(1-\pi)x''$. Then by axiom Strict FOSD, $\pi x'+(1-\pi)x'' \succ \pi x+(1-\pi)x''$, a contradiction.

(ii) Since $x \succsim x'$, $\pi'x + (1-\pi')x''$, first-order stochastically dominates $\pi'x' + (1-\pi')x''$. Thus by axiom FOSD, $\pi'x + (1-\pi')x'' \succ \pi'x' + (1-\pi')x''$. *Q.E.D.*

The theorem has a corollary for the case of expected utility:

COROLLARY *If the decision-maker satisfies axiom Preference Relation and maximizes expected utility and there exist* $x, x', x'' \in X'$ *and* $\pi\epsilon(0;1]$ *such that* $\pi x + (1-\pi)x'' \succcurlyeq \pi x' + (1-\pi)x''$*, then for all* $\pi'\epsilon(0;1]$ *:* $\pi'x + (1-\pi')x'' \succcurlyeq \pi'x' + (1-\pi')x''$.

The corollary holds since expected utility's independence axiom implies the axioms of FOSD and Strict FOSD. Note that in the thought experiment and experimental setup, the only way the recipient can learn about the decision is if the decision is implemented. d affects the recipient only via the payoff x_2^C. Thus, the theorem applies even to situations where the DM cares about not only the recipient's outcome but also about the recipient's opinion or feelings about the DM or her decision d. Thus, for consequentialist preferences, even allowing such consequences as others' opinion or the impact that the opinion has on one's self-identity, the DM's optimal split does not depend on the probability of the DM's split being implemented.

2.3. *Remarks*

Note that the axiom of Strict FOSD does not imply the axiom of FOSD. The following example gives preferences that satisfy Preference Relation and Strict FOSD but violate FOSD. The example is inspired by Machina's Mom:

> "Mom has a single indivisible item–a "treat"–which she can give to either daughter Abigail or son Benjamin. Assume that she is indifferent between Abigail getting the treat and Benjamin getting the treat, and strongly prefers either of these outcomes to the case where neither child gets it. However, in a violation of the precepts of expected utility theory, Mom strictly prefers a coin flip over either of these sure outcomes, and in particular, strictly prefers $\frac{1}{2}, \frac{1}{2}$ to any other pair of probabilities." (Machina 1989)

Machina's Mom would like to be exactly fair, thus her most preferred lottery is $(x; \frac{1}{2}, y; \frac{1}{2})$, she is indifferent between all other lotteries. Formally, for all $\pi, \pi' \in [0;1] \setminus \frac{1}{2}$: $(x; \pi, y; 1-\pi) \sim (x; \pi', y; 1-\pi')$ and $(x; \frac{1}{2}, y; \frac{1}{2}) \succ (x; \pi, y; 1-\pi)$. These preferences are complete and transitive. Axiom Strict FOSD is trivially satisfied since there is no lottery that strictly first-order stochastically dominates

another lottery. However, axiom WFOSD is violated: $(x; \frac{2}{3}, y; \frac{1}{3})$ weakly first order-stochastically dominates $(x; \frac{1}{2}, y; \frac{1}{2})$, but $(x; \frac{1}{2}, y; \frac{1}{2}) \succ (x; \frac{2}{3}, y; \frac{1}{3})$.

2.3.1. *Continuity*

In addition, continuity is not sufficient for the axiom of Strict FOSD to imply the axiom of FOSD. The preference in the previous example does not satisfy continuity, to see this note that $\{\alpha\varepsilon[0,1] : x \succsim \alpha x + (1-\alpha)y\} = [0; \frac{1}{2}) \cup (\frac{1}{2}, 1]$ and:

DEFINITION $\succsim$ is continuous if for all $p, q, r \in P$ the sets $\{\alpha \in [0,1] : \alpha p + (1-\alpha)q \succsim r\}$ and $\{\alpha \in [0,1] : r \succsim \alpha p + (1-\alpha)q\}$ are closed in [0,1].

Now consider a Machina Mom who would like to be fair, but between two unfair lotteries she prefers the one that is more fair. Formally, for all $\pi, \pi' \in [0; 1] : \pi \cdot (1-\pi) \geq \pi' \cdot (1-\pi')$ if and only if $(x; \pi, y; 1-\pi) \succsim (x; \pi', y; 1-\pi')$. The axiom of Strict FOSD is trivially satisfied since there is no lottery that strictly first-order stochastically dominates another lottery. Axiom of continuity is satisfied. However, axiom of FOSD is violated: $(x; \frac{2}{3}, y; \frac{1}{3})$ weakly first order-stochastically dominates $(x; \frac{1}{2}, y; \frac{1}{2})$, but $(x; \frac{1}{2}, y; \frac{1}{2}) \succ (x; \frac{2}{3}, y; \frac{1}{3})$.

AXIOM *(Continuity)* $\succsim$ *is continuous.*

AXIOM *(Rich domain) There are two outcomes* $x, y \in X$ *such that* $x \succ y$.

PROPOSITION *If a preference satisfies Preference Relation, Strict FOSD, Continuity, and Rich Domain then it satisfies FOSD.*

PROOF: Suppose p weakly first-order stochastically dominates q. We need to show that $p \succsim q$. Suppose not, that is $q \succ p$.

Since X is finite there exits an $\overline{x}$, $\underline{x}$ such that for all x: $\overline{x} \succsim x$, and an $x \succsim \underline{x}$. By the axiom of Rich Domain, $\overline{x} \succ \underline{x}$.

At least one of the following three cases is satisfied: (i) $\overline{x} \succ q$, (ii) $p \succ \underline{x}$ or (iii) $q \succsim \overline{x} \succ \underline{x} \succsim p$.

(i) Since p weakly first-order stochastically dominates q, and $\overline{x} \succ q$, for any $\alpha > 0$ the lottery $\alpha\overline{x} + (1-\alpha)p$ strictly first-order stochastically dominates q. But then $\{\alpha : \alpha\overline{x} + (1-\alpha)p \succsim q\} = (0, 1]$, a violation of continuity.

(ii) Since p weakly first-order stochastically dominates q, and $p \succ \underline{x}$, for any $\alpha > 0$, p strictly first-order stochastically dominates $\alpha\underline{x} + (1-\alpha)q$. But then $\{\alpha : p \succsim \alpha\underline{x} + (1-\alpha)q\} = (0, 1]$, a violation of continuity.

(iii) First we show that all elements z in the support of q satisfy $z \sim \overline{x}$. First note that by definition of $\overline{x}$, all elements in the support satisfy $\overline{x} \succsim z$. Suppose there is at least one element z such that $\overline{x} \succ z$, then $\overline{x}$ strictly first-order stochastically dominates q, which by axiom Strict FOSD implies $\overline{x} \succ q$, a contradiction. Thus, for all elements z in the support of q we have $z \sim \overline{x}$.

Second, we show that all elements z in the support of p satisfy $z \sim \underline{x}$. First note that by definition of $\underline{x}$, all elements in the support satisfy $z \succsim \underline{x}$. Suppose there is at least one element z such that $z \succ \underline{x}$, then p strictly first-order stochastically dominates $\underline{x}$, which by axiom SFOSD implies $p \succ \underline{x}$, a contradiction. Thus for all elements z in the support of p we have $z \sim \underline{x}$.

Since all elements in the support of q are indifferent to $\overline{x}$, all elements in the support of p are indifferent to $\underline{x}$, and $\overline{x}$ is strictly preferred to $\underline{x}$, q strictly first order stochastically dominates p. But that is a contradiction to p weakly first order stochastically dominating q. *Q.E.D.*

2.3.2. *Independence*

Further note that if the cardinality of the outcome space is 2, then independence is as weak an axiom as first-order stochastic dominance.

AXIOM *(Independence)* $\succsim$ *satisfies independence if for all lotteries* p, q, r *in* P : $p \succcurlyeq q \Leftrightarrow \alpha p + (1-\alpha)r \succcurlyeq \alpha q + (1-\alpha)r$.

PROPOSITION *Consider* X *with 2 elements. If* $\succsim$ *on P(X) satisfies Preference Relation, Strict FOSD and FOSD, then it satisfies Independence.*

PROOF: Without loss of generality let $X = \{x, y\}$ and $x \succcurlyeq y$. Denote $k = \alpha p + (1-\alpha)r$ and $l = \alpha q + (1-\alpha)r$.

(i) $x \sim y$

Then l weakly first-order stochastically dominates k, and vice versa. Thus by FOSD $l \succcurlyeq k$ and $k \succcurlyeq l$, thus $k \sim l$.

(ii) $x \succ y$

(ii.i) p and q are identical: Then $k = l$ and trivially $k \sim l$.

(ii.ii) $p \sim q$ but not identical: Then one must strictly first-order stochastically dominate the other, which by Strict FOSD contradicts indifference.

(ii.iii) $p \succ q$: By the lemma below, this implies $p(x) > q(x)$, and thus $p(y) < q(y)$, then k strictly first-order stochastically dominates l:

For y: $\sum_{y \succsim z} k(z) = k(y) = \alpha p(y) + (1-\alpha)r(y) < \alpha q(y) + (1-\alpha)r(y) = l(y) = \sum_{y \succsim z} l(z)$.

For x: $\sum_{x \succsim z} k(z) = 1 = \sum_{x \succsim z} l(z)$.

Thus by Strict FOSD $l \succ k$. *Q.E.D.*

LEMMA *Consider* $X = \{x, y\}$ *and* $x \succ y$. *If* $\succsim$ *on P(X) satisfies Preference Relation and Strict FOSD, then* $p \succ q$ *if and only if* $p(x) > q(x)$.

PROOF: 1.) $p \succ q$ implies $p(x) > q(x)$.

Proof by Contradiction: Suppose $p(x) \leq q(x)$.

i) $p(x) = q(x)$: This implies that $p = q$, and thus trivially by completeness $p \sim q$, a contradiction.

ii) $p(x) < q(x)$: Since $x \succ y$ this means that q strictly first order stochastically dominates p, and thus by Strict FOSD $q \succ p$, a contradiction.

2.) $p(x) > q(x)$ implies $p \succ q$: This follows from Strict FOSD. *Q.E.D.*

Note that there are examples where Independence is violated but FOSD is not. Cumulative prospect theory is one such example where the Allais paradox is allowed (thus violating Independence) but FOSD is satisfied.

2.3.3. *Sure-Thing Principle*

Savage's Sure-Thing Principle is not FOSD. If invariant, then probability does not matter. Savage gives an example of a businessman making the same choice whether or not a politician is elected; therefore, the businessman makes that choice regardless of the probability that the politician is elected.

Machina and Schmeidler (1992)'s formulation of the Sure-Thing Principle, however, is about $\frac{\partial d^*}{\partial \kappa} = 0$. There, what happens in one state of the world does not affect one's decision. FOSD makes no such predictions about $\frac{\partial d^*}{\partial \kappa} = 0$. What Machina and Schmeidler (1992) call Eventwise Monotonicity is FOSD and invariance.

2.4. *Defining consequentialism and deontic motivations*

While the previous subsections were very general in order to demonstrate an impossibility–namely, to explain variance in the probability in a consequentialist framework–we can now become less general and more concrete and apply it to an actual decision problem. We now assume that the DM has a state-independent utility function u that ranks certain outcomes and that it is twice continuously differentiable with strictly positive first derivatives with respect to the consequences (we will relax that assumption again for purely deontological preferences which will be lexicographic). Under expected utility, that u can then be chosen such that it is the DM's Bernoulli utility function. We allow the utility u of the DM to be a function of her own monetary payoff x_1, as well as the monetary payoff of the recipient x_2 to capture consequentialist other-regarding motives, and d to capture deontological motives. So the main difference to the previous subsection is that we extend the domain of the preference beyond consequences to decisions.

In the general case with all motivations present, the Bernoulli utility function satisfies $u = u(x_1, x_2, d)$. Here we can see what identifies non-consequentialist motivations. In state N, the decision d has no consequence for payoffs or for what others think or know about the DM, yet the decision does enter the utility function equally in all states of the world. This general framework now allows us to formalize the notion of decision-makers that are purely consequentialist, purely deontological, and consequentialist-deontological. Consequentialist preferences are preferences that depend on monetary payoffs and other consequences such as others' opinions of the DM.

Definition 1 Consequentialist Preferences: A preference is *consequentialist* if there exists a utility representation u such that $u = u(x)$.

We call a preference consequentialist-deontological if it incorporates concerns beyond the consequences, and considers actions or decisions that are good or bad per se:

Definition 2 Consequentialist-deontological Preferences: A preference is *consequentialist-deontological* if there exists a utility representation u such that $u = u(x, d)$.

Now let us turn to purely deontological preferences. At first, one might think they are simply the mirroring other extreme of consequentialist preferences and could thus be represented by $u = u(d)$. But, since duty is like an internal moral constraint, even fully satisfying one's duty may leave the DM with many morally permissible options rather than one unique choice. As the *Stanford Encyclopedia of Philosophy* (Alexander and Moore, 2012) puts it, "Deontological moralities, unlike most views of consequentialism, leave space for the supererogatory. A deontologist can do more that is morally praiseworthy than morality demands. A consequentialist cannot. For the consequentialist, if one's act is not morally demanded, it is morally wrong and forbidden. For the deontologist, there are acts that are neither morally wrong nor demanded." We could model duty as a moral, and thus internal constraint on the DM set of feasible decisions. Instead, we decide to model these internal constraints as the first component of a lexicographic preference. The reason we do not model duty like a budget constraint but as part of preferences, and thus lexicographic is twofold: First, unlike budget constraints, internal moral constraints are not directly observable; second, for consequentialist-deontological preferences that feature a tradeoff rather than a lexicographic ordering of these motivations, one could not model duty as an inviolable constraint. This can be formalized as a lexicographic preference, with deontological before consequentialist motivations. Note that while economists may think of our method as detecting where a DM feels most duty among competing duties, some philosophers believe there is no possibility of a genuine conflict of duties in deontological ethical theory, which can distinguish between a duty-all-other-things-being-equal (prima facie duty) and a duty-all-things-considered (categorical duty) (Alexander and Moore, 2012).

Definition 3 Deontological Preferences: A preference is called *deontological* if there exist u, f such that $u = u(d)$, and $f = f(x)$, and f.a. $(x, d), (x', d')$: $(x, d) \succsim (x', d')$ if and only if $u(d) > u(d')$ or $[u(d) = u(d')$ and $f(x) \geqq f(x')]$.

Observable choice behavior then allows us to experimentally identify whether subjects have preferences where both motivations are present (i.e., whether their preferences belong to the category of consequentialist-deontological preferences). In particular, we will ask how exogenous variation in the probability π of the decision being consequential impacts the optimal decision. Note that the DM has one choice variable only, d, but by varying the probability of her decision being consequential we can identify whether she cares only about the consequences or also about the decision per se. Since she has only one choice variable it is often useful to consider her indirect objective function $V(d)$.

2.5. *Consequentialists who maximize Expected Utility*

Given expected utility, the DM maximizes:

$$E[u(x,d)] = \pi u(x_1^C, x_2^C, d) + (1\text{-}\pi)u(x_1^N, x_2^N, d)$$

and her indirect objective function in case of the dictator game can be written as:

$$V(d) = \pi u(\omega - d, d, d) + (1-\pi)u(\omega - \kappa, \kappa, d).$$

Limiting attention to pure consequentialists the problem simplifies to:

$$E[u(x)] = \pi u(x_1^C, x_2^C) + (1-\pi)u(x_1^N, x_2^N)$$

and the indirect objective function to:

$$V(d) = \pi u(\omega - d, d) + (1-\pi)u(\omega - \kappa, \kappa).$$

Note that now the d does not enter in the second term, which corresponds to state N.

Let us first consider the simplest example of a consequentialist preference, homo oeconomicus:

EXAMPLE 1 **(Homo oeconomicus)** Homo oeconomicus is a consequentialist whose preferences depend only on her own outcome. Her preference can be represented by a Bernoulli utility function with $u = u(x_1)$. Her constrained maximization problem is thus $max_d V(d) = \pi u(\omega - d) + (1 - \pi)u(\omega - \kappa)$ subject to $0 \leq d \leq \omega$. As the objective function is proportional to $u(\omega - d)$, the unique maximizer is $d^* = 0$. Observe that the optimal decision d^* of homo oeconomicus does not depend on the probability π of the decision being consequential.

Intuitively, as the probability of the sharing decision being implemented varies, both its benefits and costs vary in the same way. Is this independence of the optimal decision d^* true for consequentialist preferences more generally? Let us investigate this with another example of a consequentialist preference, Fehr-Schmidt:

EXAMPLE 2 **(Fehr-Schmidt preferences)** Fehr and Schmidt (1999) propose preferences such that the DM is a consequentialist who cares about her own and others' monetary payoffs. The idea is that decision-makers dislike inequality, but dislike inequality in their disfavor even more. Recall that the Fehr-Schmidt utility function is $u(x) = x_1 - \alpha max\{x_2 - x_1, 0\} - \beta max\{x_1 - x_2, 0\}$, where $\beta < \alpha$ and $0 \leq \beta < 1$. We can write the decision-maker's expected utility as:[7]

[7]The functional form $x_1^C - \alpha max\{x_2^C - x_1^C, 0\} - \beta max\{x_1^C - x_2^C, 0\}$ comes directly from Fehr and Schmidt (1999).

$$\begin{aligned}E[u(x)] =&\pi\left(x_1^C - \alpha max\{x_2^C - x_1^C, 0\} - \beta max\{x_1^C - x_2^C, 0\}\right)\\ &+(1-\pi)\left(x_1^N - \alpha max\{x_2^N - x_1^N, 0\} - \beta max\{x_1^N - x_2^N, 0\}\right)\end{aligned}$$

The indirect objective function is then:

$$\begin{aligned}V(d) =&\pi\left(\omega - d - \alpha max\{2d-\omega, 0\} - \beta max\{\omega - 2d, 0\}\right)\\ &+(1-\pi)\left(\omega - \kappa - \alpha max\{2\kappa-\omega, 0\} - \beta max\{\omega - 2\kappa, 0\}\right)\end{aligned}$$

V obtains a maximum wherever the first summand does. Thus, as usual a Fehr-Schmidt decision-maker will choose $d = 0$ if $\beta < \frac{1}{2}$, and $d = \frac{\omega}{2}$ if $\beta > \frac{1}{2}$, and for $\beta = \frac{1}{2}$, she is indifferent between all donations that are no more than half the endowment. The optimal donation does not depend on the probability.

Another famous example of social preferences are Andreoni preferences (Andreoni 1990):

EXAMPLE 3 (**Andreoni preferences**) Andreoni (1990) points out that DMs in a public goods contribution framework empirically seem to derive utility not only from the total amount of the public good G provided, but also from her contribution g. Deontological motivations do not necessarily coincide with warm glow. In the Andreoni framework one cannot tell the consequential and the deontological apart. It is a theory of individuals caring about their actions, potentially for reasons of duty but also for reasons of social audience (Andreoni and Bernheim 2009). Both interpretations are consistent with the formal model, but the verbal description of "impure altruism" suggests a consequentialist understanding, that is, G is the public good consequences of the decision and g includes the social reactions to the generosity of the individual decision. Thus, a DM with warm-glow preferences is a consequentialist whose preferences depend on her own outcome, the charitable recipient's outcomes, and the decision that is implemented. In the Andreoni-framework, it is not possible that a DM decides to contribute g but then her decision is not carried out. Assume this can happen and that in this case she contributes some constant κ (think of it as zero). All others contribute G_{-DM} to the public good in every state of the world. Then we can write the decision-maker's expected utility as:[8]

$$E[u(x_1, g, G)] = \pi u(x_1^C, g^C, G^C) + (1-\pi)u(x_1^N, g^N, G^N)$$

The indirect objective function is then:

$$V(d) = \pi u(\omega - d, d, G_{-DM} + d) + (1-\pi)u(\omega - \kappa, \kappa, G_{-DM} + \kappa)$$

[8]The function, $u(x_1^C, g^C, G^C)$, comes directly from Andreoni (1990).

Note that the objective function is affine in $u(\omega - d, d, G_{-DM} + d)$. Thus d^* does not depend on π.

A recent example of social preferences are Benabou-Tirole preferences (Bénabou and Tirole 2011):

EXAMPLE 4 Bénabou and Tirole (2011) models moral decision-making as part of identity investment that prevents future deviant behavior. Deontological motivations do not necessarily coincide with this. The 5th commandment in the Bible reads "You shall not kill," not "You shall minimize killing", the latter being consequentialist. In the Benabou-Tirole framework, one also cannot tell the consequential and the deontological apart. Individuals concerned about identity may care about decisions, g^N, or the decisions carried out, g^C. Both interpretations are consistent with their model. The verbal description of "identity investment" suggests a consequentialist interpretation. However, even if not, our empirical investigations can be interpreted as providing revealed preference evidence–rather than survey evidence–for the "taboo thoughts" cited in Bénabou and Tirole (2011).

FACT ***(consequentialist EU maximizers)*** *For a consequentialist DM who satisfies the assumptions of expected utility theory the optimal d^* does not depend on π.*

PROOF: A consequentialist who satisfies the axioms of expected utility theory faces the constrained optimization problem $V(d) = \pi u(\omega - d, d) + (1 - \pi)u(\omega - \kappa, \kappa)$ subject to $0 \leq d \leq \omega$. Note that the indirect objective function V is affine in $u(\omega - d, d)$. Thus, the optimal d does not depend on π.[9] *Q.E.D.*

Another way to see this is to look at the first-order condition, which can be written as $\frac{u_1(\omega-d,d)}{u_2(\omega-d,d)} = 1$, so the marginal rate of substition equals the marginal cost of donation of 1. Therefore, under expected utility, for any consequentialist DM the amount shared does not vary in the probability.

2.6. *Purely deontological preferences*

We say that the DM has deontological motivations if her Bernoulli utility does not only depend on the consequences but also on the decision itself. The DM cares about her decision even if it is without any monetary or non-monetary consequence for herself and others. Thus, even if the decision is never implemented, no one ever learns about which decision she took, and thus no one's opinion about the DM changes as a consequence of her decision, but she still cares about the decision.

THEOREM 2 ***(Deontological preferences)*** *For purely deontological preferences the optimal decision d^* is constant in the probability π.*

It is a natural question to ask if deontological moral philosophy even applies to situations under (objective) uncertainty. A first response is that the natural world is always uncertain, so any moral philosophy that aims to provide guidance to people in the natural world presumably must apply

[9]Note that there could be more than one optimal d, but then the solution set does not depend on π.

to decision-making under uncertainty. Let us illustrate this point with what Kant thinks about uncertainty in the famous ax-murderer example: "Es ist doch möglich, daß, nachdem du dem Mörder auf die Frage, ob der von ihm Angefeindete zu Hause sei, ehrlicherweise mit ja geantwortet hast, dieser doch unbemerkt ausgegangen ist und so dem Mörder nicht in den Wurf gekommen, die That also nicht geschehen wäre" (Kant 1797). He states that there is always some uncertainty about the consequences of saying the truth to the ax-murderer, so therefore, one should do one's duty to say the truth regardless of what happens to the ax-murderer or the victim (i.e., the victim happens to have left the house unnoticed by you).

This is because in these lexicographic preferences, a person is either pure deontological or pure consequentialist in comparing possible decisions. Formally, there is no trade-off. A lexicographic deontologist maximizes $u(d)$ first, then there is a compact set where she maximizes $v(x)$ next. Our theorem applies to either the pure consequentialist portion $v(x)$ or the deontological portion $u(d)$.

2.7. *Consequentialist-deontological preferences*

Theorems 1 and 2 show that neither consequentialist nor purely deontological preferences predict behavioral changes as the probability of being consequential changes. Now we give a simple example of consequentialist-deontological preferences where the optimal decision changes as the probability of being consequentialist changes. To that end, we consider an additive utility functions that depends on the decision d and on only one consequence, the payoff for the DM herself:

EXAMPLE 5 $u = u(x_1, d) = x_1 + b(d)$, where $b_1 > 0$ and $b_{11} < 0$.

Then $V(d) = \pi(\omega - d) + (1 - \pi)(\omega - \kappa) + b(d)$ is strictly concave in d. The first-order condition is $b_1(d) = \pi$ and thus for an interior solution $\frac{\partial d^*}{\partial \pi} = \frac{1}{b_{11}(d)} < 0$.

Thus in the above example d^* is decreasing in the probability of being consequential. This result means that the lower the probability that the DM's decision is implemented, the more she donates. At first glance, this result may seem somewhat counter-intuitive, but it is consistent with arguments made by political conservatives that wealthy people vote for generous redistribution when their probability of being pivotal is low; while in real-life, where their decision is definitely implemented, they may donate only a little. This is also consistent with observational evidence on the decision to sign up to be a bone marrow donor (Bergstrom et al. 2009) and the decision to abort a fetus with Down Syndrome (Choi et al. 2012). *In our setup, the benefit of altruism is always there, but the costs are only incurred with probability* $1 - \pi$. So, the lower the probability the decision is executed, the lower the cost of making the decision, and thus we should expect to see *more* altruism.[10]

For a slightly more general model: let $u(x_1, d) = f(x_1) + b(d)$. Then, $U(x_1, d) = \pi(f(x_1^C) + b(d)) + (1 - \pi)(f(x_1^N) + b(d))$ and $V(d) = \pi f(\omega - d) + (1 - \pi)f(\omega - \kappa) + b(d)$. The first order condition is: $\frac{\partial V(d)}{\partial d} = -\pi f_1(\omega - d) + b_1(d) = 0$. For d^* to be a maximum, then $\frac{\partial^2 V(d)}{\partial d^2} = \pi f_{11}(\omega - d) + b_{11}(d) < 0$. By the implicit function theorem, $\frac{\partial d^*}{\partial \pi} = \frac{f_1(\omega - d^*)}{\pi f_{11}(\omega - d^*) + b_{11}(d^*)} < 0$, since utility is increasing in its own outcomes and the denominator, which is the second derivative of the indirect objective function, is

[10] Utility in money does not have to be linear to obtain this result.

negative. Note that the recipient's payoff is a function of the DM's payoffs, but as long as other-regarding concerns are concave then the sum of utility from its own payoffs and utility from others' payoffs is still concave and the above result holds. Decisions do not have to be continuous to obtain this result. If decisions are discrete, then the behavior of a mixed consequentialist-deontological person is jumpy (i.e., it weakly increases as her decision becomes less consequential). Note that if the consequentialist and deontological choice is the same, then the choice is still invariant to the implementation probability: $f_1(\omega - d) = b_1(d) = 0$, then $\frac{\partial d^*}{\partial \pi} = 0$.

EXAMPLE 6 **(Impure altruism and deontological motivations)** In example 3 we showed that Andreoni-preferences for warm-glow are purely consequential. Now let us extend the Andreoni-preferences to allow for deontological motivations and assume a utility function of the form $u = u(x_1, g, G, d)$. Thus the consequences x are now the triple $x = (x_1, g, G)$. The DM decides how much to contribute, that is chooses a d (where d is affordable $0 \leq d \leq \omega$). In the consequential state the decision gets implemented and thus $g^C = d$, whereas in the nonconsequential state $g^N = \kappa$.

$$E[u(x_1, g, G, d)] = \pi u(x_1^C, g^C, G^C, d) + (1 - \pi)u(x_1^N, g^N, G^N, d)$$

The indirect objective function is then:

$$V(d) = \pi u(\omega - d, d, G_{-DM} + d, d) + (1 - \pi)u(\omega - \kappa, \kappa, G_{-DM} + \kappa, d)$$

Now d^* can vary in π.

Non-Additive Utility

For more complicated utility functions, non-additive or non-globally convex ones, it is possible to generate examples where $\frac{\partial d^*}{\partial \pi} = \frac{1}{b_{11}(d)} > 0$. Suppose the DM has preferences represented by $u = u(x_1, d)$. Assume that the first derivatives are positive (monotonicity), and that $u_{11} < 0$ and $u_{22} < 0$ (risk-aversion). Then the DM maximizes $V(d) = \pi u(\omega - d, d) + (1 - \pi)u(\omega - \kappa, d)$. The first order condition is $-\pi u_1(\omega - d, d) + \pi u_2(\omega - d, d) + (1 - \pi)u_2(\omega - \kappa, d) = 0$. By the implicit function theorem, and simplifying using the first order condition gives:

$$\frac{\partial d^*}{\partial \pi} = \frac{1}{\pi^2}[-2u_{12}(\omega - d, d) + u_{11}(\omega - d, d) + u_{22}(\omega - d, d) + \frac{1 - \pi}{\pi}u_{22}(\omega - \kappa, d)]^{-1}u_2(\omega - \kappa, d)$$

So for sufficiently negative $u_{12}(\omega - d, d)$ we can get $\frac{\partial d^*}{\partial \pi} > 0$. Utility functions that are not globally convex can lead to local maxima that, when the decision is less consequential, can lead to jumps to maxima involving lower d.

EXAMPLE 7 **(Bliss Point)** $u(x_1, x_2, d) = (1 - \mu)\Big(-(1 - \lambda)(\omega - x_1)^2 - \lambda(\omega - x_2)^2\Big) - \mu(\delta - d)^2$, where $0 \leq \delta \leq \omega$ and $0 \leq \mu, \lambda \leq 1$. In our thought experiment, $V(d) = \pi(1 - \mu)\Big(-(1 - \lambda)d^2 - \lambda(\omega - d)^2\Big)$ $+ (1 - \pi)(1 - \mu)\Big(-(1 - \lambda)\kappa^2 - \lambda(\omega - \kappa)^2\Big) - \mu(\delta - d)^2$. For a DM who is pure consequentialist

($\mu = 0$), the function obtains its global maxima at a blisspoint: $d^* = \lambda\omega \equiv d_c^*$. For a DM who is pure deontological ($\mu = 1$) there exists a blisspoint $d^* = \delta \equiv d_d^*$. We now look at a person with mixed preferences. There is a unique critical point where the function obtains its global maxima: $d^* = \frac{\pi(1-\mu)}{\pi(1-\mu)+\mu}\lambda\omega + \frac{\mu}{\pi(1-\mu)+\mu}\delta = \frac{\pi(1-\mu)}{\pi(1-\mu)+\mu}d_c^* + \frac{\mu}{\pi(1-\mu)+\mu}d_d^*$. As you can see, d^* is a weighted mean of the two bliss points and if $d_c^* \neq d_d^*$ it depends on π: $\frac{\partial d^*}{\partial \pi} = \frac{(1-\mu)\mu\left(d_c^* - d_d^*\right)}{(\pi(1-\mu)+\mu)^2}$. The relation between d_c^* and d_d^* determine the sign of this expression. If $d_c^* > d_d^*$, so the bliss point for a consequentialist is to the right of the bliss point for a deontologist, then as the probability of being consequential increases, the d^* increases as well. Such a situation can arise, for example, if social audience concerns are strong and the duty to donate to others is weak, perhaps because the duty to one's own family is strong. Bell-shaped utility functions commonly used in estimates of policy choices by politicians lead to such a scenario.

This formulation also addresses the possibility of competing duties, such as the duty to charity versus the duty to family or self. Reducing the probability a decision is implemented should lead to decisions that align more with the direction where one feels the greatest duty. The direction of the decision changes gives insight into these competing duties, and the location of the optimand for one's greatest duty.

2.8. *Other explanations*

2.8.1. *Ex-ante fairness*

A potential confound to our explanation of deontological motivations is that people could have preferences over the lotteries themselves if they view them as procedures, rather than if their preferences are fundamentally driven by the prizes (consequences or the decision). Formally, this is a violation of first-order stochastic dominance, and as such might be viewed as implausible, but a famous example articulated by Machina (1989) and recapitulated in Sub-section 2.3 shows how this might not be as implausible as first thought. In our experimental setup, for example a subject might target the expected income of the recipient, and thus vary the decision in the probability.

EXAMPLE 8 Targeting the recipient's expected income. Consider the following preferences $U(x_1, x_2) = E[x_1] + a(E[x_2]) = \pi x_1^C + (1-\pi)x_1^N + a\left(\pi x_2^C + (1-\pi)x_2^N\right)$. Let a be a function that captures altruism and let it be strictly increasing and strictly concave. Note that this objective function is not linear in the probabilities. The indirect objective function is $V(d) = \pi(\omega - d) + (1-\pi)(\omega - \kappa) + a(\pi d + (1-\pi)\kappa)$. The first-order condition is $a_1(\pi d + (1-\pi)\kappa) = 1$. By the implicit function theorem, $\frac{\partial d^*}{\partial \pi} = \frac{\kappa - d^*}{\pi}$. Thus the optimal decision changes in the probability. In two special cases, it is easy to determine the sign of the derivative, even if d^* itself is not (yet) known: if $\kappa = 0$, then $\frac{\partial d^*}{\partial \pi} \leqq 0$, and if $\kappa = \omega$, then $\frac{\partial d^*}{\partial \pi} \geqq 0$.

Let us look at a more general case: $U = f(E[u(x_1)], E[\widetilde{u}(x_2)])$, where f is $f_1, f_2 > 0$ (strictly increasing), $f_{12}f_1f_2 - f_{11}f_2^2 - f_{22}f_1^2 > 0$ (strictly quasi-concave), ($f_{12}f_2 - f_{22}f_1 > 0$ and $f_{12}f_1 - f_{11}f_2 \geq 0$) or ($f_{12}f_2 - f_{22}f_1 \geq 0$ and $f_{12}f_1 - f_{11}f_2 > 0$) (strictly normal in in one argument, weakly

normal in the other), $u, \widetilde{u}$ is $u_1, \widetilde{u}_1 > 0$ (strictly increasing), $u_{11}, \widetilde{u}_{11} \leq 0$ (weakly concave) and $\pi > 0$. Then, the indirect objective function is

$$V(d) = f\left(\pi u(\omega - d) + (1-\pi)\, u(\omega - \kappa), \pi \widetilde{u}(d) + (1-\pi)\, \widetilde{u}(\kappa)\right)$$

Note that $V(d)$ is globally strongly concave:

$$\begin{aligned}\frac{1}{\pi}\frac{\partial^2 V(d)}{(\partial d)^2} = &-\left(2 f_{12} f_1 f_2 - f_{11} f_2^2 - f_{22} f_1^2\right) \frac{1}{f_2^2} \pi u_1^2 (\omega - d) \\ &+ f_1 u_{11}(\omega - d) + f_2 \widetilde{u}_{11}(d) < 0\end{aligned}$$

So, there exists a unique solution. The First-order condition for this problem is $\frac{\widetilde{u}_1(d)}{u_1(\omega - d)} - \frac{f_1}{f_2} = 0 \equiv F$. The FOC defines d^* implicitly as a function of π. By the implicit function theorem $\frac{\partial d^*}{\partial \pi} = -\frac{\frac{\partial F(d^*,\pi)}{\partial \pi}}{\frac{\partial F(d^*,\pi)}{\partial d^*}}$. As $\frac{\partial F(d^*,\pi)}{\partial d^*}$ has sign of $\frac{\partial^2 V(d)}{(\partial d)^2} < 0$: $sgn\left(\frac{\partial d^*}{\partial \pi}\right) = sgn\left(\frac{\partial F(d^*,\pi)}{\partial \pi}\right)$. It can be shown that:

$$\begin{aligned}\frac{\partial F(d^*,\pi)}{\partial \pi} = &\frac{\widetilde{u}_1(d^*)}{f_1}\left(f_{12} f_1 - f_{11} f_2\right)\left[u(\omega - d^*) - u(\omega - \kappa)\right] \\ &+ \frac{u_1(\omega - d^*)}{f_2}\left(f_{12} f_2 - f_{22} f_1\right)\left[\widetilde{u}(\kappa) - \widetilde{u}(d^*)\right]\end{aligned}$$

So the sign of $\frac{\partial d^*}{\partial \pi}(\pi)$ depends on the difference between $d^*(\pi)$ and κ:

For $d^*(\pi) = \kappa$: $\frac{\partial F(d^*,\pi)}{\partial \pi} = 0$ thus $\frac{\partial d^*}{\partial \pi}(\pi) = 0$

For $d^*(\pi) < \kappa$: $\frac{\partial F(d^*,\pi)}{\partial \pi} > 0$ thus $\frac{\partial d^*}{\partial \pi}(\pi) > 0$

For $d^*(\pi) > \kappa$: $\frac{\partial F(d^*,\pi)}{\partial \pi} < 0$ thus $\frac{\partial d^*}{\partial \pi}(\pi) < 0$

Now if $\kappa = 0$, then $\frac{\partial d^*}{\partial \pi} \leqslant 0$, while for $\kappa = \omega$ $\frac{\partial d^*}{\partial \pi} \geqslant 0$. Thus experimentally, by varying κ we can test whether people have these ex-ante considerations. Note that $\frac{\partial d^*}{\partial \kappa} = 0$, which we also examine empirically.

2.8.2. *Cognition costs*

Another possible explanation for variance in the probability might be cognition costs. Cognition costs are a consequence, but unlike the other consequences, they are not captured in our consequentialist framework since they are incurred during the decision and are a consequence that even arises if the nonconsequential state is realized. To fix ideas, consider the following model: $u = u(x_1, x_2, \gamma)$, where $u_1, u_2 > 0$, $u_\gamma < 0$ and $\gamma \geq 0$. In addition, let us assume that utility is continious. The DM can compute the optimal decision, but to do so, she incurs a cognition cost $\gamma > 0$, otherwise she can make a heuristic (fixed) choice $\bar{d}$ for which (normalized) costs are 0. We have no model of what the heuristic choice is, and in principle it could be anything, but recent experimental work argues that the heuristic choice tends to be a cooperative or fair one (Rand et al. 2012) so, for example, the reader might think of $\bar{d} = \frac{\omega}{2}$. In any case, expected utility from the heuristic

choice is $V(\bar{d}) = \pi u(\omega - \bar{d}, \bar{d}, 0) + (1-\pi)u(\omega - \kappa, \kappa, 0)$. By constrast, for a non-heuristic choice, $V(d) = \pi u(\omega - d, d, \gamma) + (1-\pi)u(\omega - \kappa, \kappa, \gamma)$. Define $\check{d} \equiv argmax V(d)$. Obviously, $\check{d}$ does not vary in π. The DM will choose to act heuristically if $V(\check{d}) < V(\bar{d})$ or

$$\begin{aligned} F(\pi) \equiv V(\check{d}) - V(\bar{d}) = \pi \left(u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0)\right) \\ + (1-\pi)\left(u(\omega - \kappa, \kappa, \gamma) - u(\omega - \kappa, \kappa, 0)\right) < 0 \end{aligned}$$

Since $(1-\pi)\left(u(\omega - \kappa, \kappa, \gamma) - u(\omega - \kappa, \kappa, 0)\right) < 0$, we can distinguish two cases:

i) If $u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0) < 0$, $F(\pi)$ is always negative, so the person uses the heurisic choice, independent of π.

ii) In the other case, $u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0) > 0$, there exists a unique $\widetilde{\pi}$ with $0 < \widetilde{\pi} < 1$ such that $F(\widetilde{\pi}) = 0$, the person switches from heuristic to non heuristic. This derives from the fact that in this case $F(\pi)$ is strictly monotone in π, $F(0) < 0$ and $F(1) > 0$, so for probabilities of being consequential close to 1 computing is better, and for probabilities close to zero, the heuristic is better. Since $\check{d} \neq \bar{d}$, this means that such cognition costs predict that even a consequentialist DM will not be invariant to the probability. For the rest of this section, we will focus on this case.

Now suppose we vary the cognition cost, that is, we do an exercise in comparative statics and investigate how $\widetilde{\pi}$ varies in γ, and note that

$$\frac{\partial \widetilde{\pi}}{\partial \gamma} = \frac{-\widetilde{\pi} u_3(\omega - \check{d}, \check{d}, \gamma) - (1-\widetilde{\pi})\, u_3(\omega - \kappa, \kappa, \gamma)}{u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0) + u(\omega - \kappa, \kappa, 0) - u(\omega - \kappa, \kappa, \gamma)} > 0,$$

that is, the higher the cognition costs, the higher the threshold for probability being consequential such that computation is the better choice. Obviously, there are some very low γ and some very high γ such that locally, $\widetilde{\pi}$ is a constant function of γ, but there, the above assumptions are violated. The figure shows when, as a function of a probability, someone would incur a given cognition cost. So if we could experimentally vary not only probability but also cognition costs and then observe it, the cognition cost story predicts the pattern shown in the figure.

In summary, variation in the decision d with respect to π is consistent with decision-makers switching to a heuristic $\bar{d}$, which may be higher or lower than the preferred choice $\check{d}$, leading to the inability to infer consequentialist-deontological preferences. If decision-makers have different γ or different $\bar{d}$, then we might observe a smooth $\frac{\delta d}{\delta \pi}$. A cognition-costs model, however, would predict that time spent on the survey also changes as d changes with π. Our research design provides a second test of the cognition-costs model: Subjects with greater cognition costs should have $\frac{\delta d}{\delta \pi} = 0$ for a larger range of π near 0. An S-shape curve in cognition costs incurred and thus in decisions with respect to π, is more shifted, the higher cognition costs are. This formal modeling and experimental test of cognition costs seems to be rare in the literature. For a previous example, albeit one that does not have the decision-maker solve the metaproblem optimally, see Wilcox (1993). Figure 3 plots the cognition costs incurred against π. The dotted line is for the subject experiencing low cognition costs while the dashed line is for the subject experiencing high cognition costs.

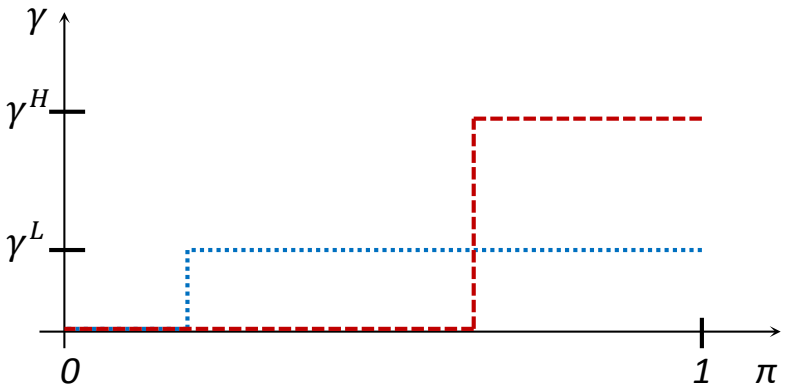


FIGURE 3.—

3. EXPERIMENTAL METHODOLOGY

In our lab experiments, we operationalize our thought experiment by asking subjects to make a donation decision with the knowledge that we will shred their decision when it is not implemented. Participants first see a demonstration of a public randomization device (Wheel of Fortune) and a paper shredder; the shredding bin is opened to publicly verify that materials will truly be destroyed.[11] The donation recipient was Doctors Without Borders in sessions held in German-speaking countries to maximize the likelihood that the duty to donate would be high. We ran the lab study in Zurich, Hamburg, and Magdeburg.

In Zurich, subjects are asked three IQ tasks first to ensure they take the study seriously. If at least one answer is correct, they proceed to the donation decision. Subjects are randomly assigned to low ($\pi = \frac{3}{16}$) or high probability ($\pi = \frac{15}{16}$) of implementation and to minimum ($\kappa = 0$) or maximum ($\kappa = \omega$) donation in the non-consequential state. After the Wheel of Fortune is spun, envelopes that are to be destroyed are collected and shredded. The remainder are opened and participants are paid. Mean and median age of subjects was 23.[12] One session was held in the classroom, where the endowment was 10Chf instead of 20Chf, so all our results are reported in terms of percentages.[13] Among 264 subjects, 71 envelopes were opened.

Our second set of lab experiments paired the donation decision with a modified moral trolley vignette. Individuals are asked whether they will kill one person to save many.[14] We vary the number of people who would be saved in order to correlate decisions indicating mixed consequentialist and deontological motivations in the vignette task with decisions in the donation task. We conduct this experiment in Hamburg and Magdeburg. We set the low probability to ($\pi = \frac{1}{16}$) so the instruction

[11] Shredding may be referred to as, the unstrategic method, because fewer than one datapoint per subject is collected, in contrast to the strategy method, which does the opposite.

[12] We dropped extreme outliers in terms of age as they are likely to be outliers among most dimensions. People under 18 were not allowed to participate. People 30 or older were also excluded.

[13] We had difficulty recruiting and faced high expenses, so we moved our lab experiment to Germany.

[14] In a separate sample without the donation decision, we piloted the moral trolley vignette to observe which one would be the most elastic to changes in the number of people saved.

pages could show the same number of numbers from the randomization device. Doing so greatly increasing our experimental costs. Among 975 subjects, 173 envelopes were opened. We also removed the IQ test added the moral trolley questionnaire.

The third set of experiments estimates structurally the trade-off between consequentialist and deontological motivations. To do so, we greatly increase our sample size by recruiting on Amazon Mechanical Turk, a labor market intermediary (LMI). The LMI can be used to implement anything from a natural field experiment to a laboratory experiment. Workers come to the marketplace naturally and are unaware they are in an experiment at the time of arrival, and this lack of awareness alleviates Hawthorne effects (Orne 1962; Titchener 1967). Through an interface provided by the LMI, registered users perform tasks posted by buyers for money. The tasks are generally simple for humans to complete, but difficult for computers. Common tasks include captioning photographs, extracting data from scanned documents, and transcribing audio clips.

To lock workers in and prevent attrition of different types of people in response to treatment, we first asked them to do data transcription of three paragraphs of text (Chen 2012; Chen and Horton 2014). This task was sufficiently tedious that no one was likely to do it "for fun," and it was sufficiently simple that all participants could do the task. The source text was machine-translated to prevent subjects from finding the text elsewhere on the Internet. A paragraph takes about 100 seconds to enter so a payment of 10 cents per paragraph is equivalent to $28.80 per 8-hour day. The current federal minimum wage in the Unites States is $58/day.

After the lock-in task, subjects have an opportunity to split their bonus with the charitable recipient, the Red Cross. Workers then provided their gender, age, country of residence, religion, and how often they attend religious services. After work was completed and according to the original expiration date listed on the LMI, bonuses were calculated and workers were notified of their earnings. We do not shred decisions and are able to collect data for very low implementation probabilities. We had 902 decisions from 902 subjects.

We ran two MTurk experiments. In both experiments, participants are randomly assigned to one of five groups that differ in π, the probability of the decision being consequential: 100%, 66%, 33%, 5%, and 1%. All subjects are in the role of dictator and the recipient is the Red Cross.[15] All subjects are told about the implementation probability. In one experiment, we additionally randomize κ to be 50 cents (maximum) and 0 cents (minimum) in the different treatment arms. In a second experiment, we randomize κ to be 10 cents or unknown to workers (they are told the computer is making a determination) and we draw κ from a uniform distribution. When κ was unknown, we also asked workers what they believed would be the amount donated if the computer made the decision. We present both the raw data as well as regression specifications that include indicator variables for κ.

The results do not hinge on using the κ-unknown treatment arm. We randomize κ to be 10 cents or unknown in order to assess the confound that is raised by the experimenter observing the

[15]Most subjects are from the U.S. and India, and we believed the Red Cross to be more well-known in these countries.

data when the decision is not implemented. Because decisions are not shredded, another reason for subjects to vary their behavior with respect to the probability of implementation is the reputational cost of the decision existing in both states of the world. The reputational cost and the deontological benefit from the generous decision cannot be distinguished when the experimenter observes. However, dictators concerned about reputation may justify their selfishness with nature's choice (Andreoni and Bernheim 2009) so we examine how frequently dictators choose nature's choice when κ is 10 cents or unknown. 18% of subjects gave 10 cents in the "10 Cents" treatment while 14% gave 10 cents in the "Unknown" treatment. This difference is not statistically significant. Thus, the social audience mechanism driving the results in Andreoni and Bernheim (2009) do not appear relevant here.

3.1. *Specification*

The empirical specification examines the effect of treatment on donation:

$$Donation_i = \beta_0 + \beta_1 Treatment_i + \beta_2 X_i + \varepsilon_i \tag{1}$$

$Treatment_i$ represents the treatment group for individual i (sometimes represented as π, the probability a decision is consequential) and X_i represents individual demographic characteristics. We display the raw data means, distributions, and results from the Wilcoxon-Mann-Whitney test for differences in distributions of donations, and ordinary least squares regressions. When we included covariates, country of origin was coded as United States and India with the omitted category as other; religion was coded as Christian, Hindu, and Atheist with the omitted category as other; religious services attendance was coded as never, once a year, once a month, once a week, or multiple times a week.[16]

We next estimate how sensitive the decision d is to π for each individual as predicted from their demographic characteristics. In essence, we construct synthetic cohorts to emulate a within-subject design. Formally, we estimate:

$$Donation_i = \beta_0 \pi_i + \beta_1 \mathbf{X}_i \pi_i + \alpha \mathbf{X}_i + \varepsilon_i$$

We interpret the change in d to π as measuring the mixed consequentialist-deontological motives. Intuitively, if $\mathbf{X}_i$ were country-fixed effects, this would be like computing country-level averages of $\frac{\delta d}{\delta \pi}$. Each demographic variable contributes to the effect of probability of being consequential on the donation.

We then compute for each individual:

$$MixedConsequentialistDeontological_i = |\hat{\beta}_0 + \widehat{\beta}_1 \mathbf{X}_i|$$

We use all the demographic characteristics to construct a mixed consequentialist-deontological

[16]Some regressions also code for levels of respect for their parents, police, and their boss, respectively: not at all, not much, some, a little, and a lot.

score. Each demographic variable contributes to the effect of probability of being consequential on the donation. Each subject's demographic variables are then used to calculate a predicted mixed consequentialist-deontological score by taking the absolute value of the sum of the contributions of their demographic characteristics along with the constant term. We interpret the change in d to π as measuring mixed consequentialist-deontological motives. Intuitively, if $\mathbf{X}_i$ were a dummy indicator for being male, this would be like computing $\frac{\delta d}{\delta \pi}$ for the average male. Males may be less generous than females, but generosity of both males and females may decrease with π. Whether $\frac{\delta d}{\delta \pi} < 0$ in different sub-populations allows investigation of the possibility that people's duties differ.

3.2. *Power Calculation to Determine Optimal Treatment Ratio*

We conduct a power calculation to determine the optimal ratio of treatment to control subjects. For example, when our two probabilities are 15/16 and 3/16, the data collection for low π is five times more expensive. Our estimand is: $\hat{ß} = E(T) - E(C) = \frac{\sum T}{n_T} - \frac{\sum C}{n_C}$. We seek to minimize $Var(\hat{ß}) = \frac{\hat{\sigma_T^2}}{n_T} - \frac{\hat{\sigma_C^2}}{n_C}$ subject to the budget constraint that $n_T c_T + n_C c_C \leq I$. The first-order conditions of the Lagrangian are $-\frac{1}{n_T^2} = -\lambda c_T$ and $-\frac{\gamma}{n_C^2} = -\lambda c_C$ where $\gamma = \frac{\hat{\sigma_C^2}}{\hat{\sigma_T^2}}$. This determines the optimal ratio of data collection to be: $\frac{n_T^2}{n_C^2} = \gamma \frac{c_C}{c_T}$. Intuitively, as the cost of data collection for treatment increases, we collect more control. As the variance of the treatment sample increases, we collect more treatment. Sample variance among low π subjects was higher in the pilot, which required a roughly 1:1 ratio of opened envelopes.

3.3. *Related Research Designs*

Our experimental design can be contrasted with Andreoni and Bernheim (2009), which is also a modified dictator game with random implementation probabilities. First, in their study, the decision is not shredded so the dictator could be motivated by what the experimenter infers (as they acknowledge). The reputational cost of the decision is present in both states of the world, so the reputational cost and deontological benefit are not distinguishable. Second, in our study, both the probability and the realization are public. In their study, dictators can hide their selfishness behind nature's choice, which is not observed by recipients. Third, in our study, the recipient is even not present to further reduce social audience concerns. Finally, the results differ: In their study, dictators become *more* generous as the probability of implementation increases, while in our study, dictators become *less* generous as the implementation probability increases. Random implementation probabilities can also be contrasted with a large literature in psychology that varies the probability that one's help will have an impact (Batson et al. 1991; Smith et al. 1989). These studies examine whether one's help *actually helps*, rather than whether one's help *will be carried out*, an important distinction, since the cost of the decision is experienced by subjects whether or not their decision to help is effective.

Two recent observational studies (Bergstrom et al. 2009; Choi et al. 2012) mimic our experimental design and are suggestive that deontological motivations are present outside the lab. African-

Americans in the U.S. are less likely to register for bone marrow donation while Caucasian-Americans are more likely to register for bone marrow donations (Bergstrom et al. 2009). However, conditional on registration, African-Americans are more likely to be asked to donate than Caucasian-Americans. Our model suggests that ethnicities that have a low probability of being called to donate bone marrow are going to be more altruistic in signing up for bone marrow donation. For a different context, Choi et al. (2012) reports that as women's decisions to abort a fetus with Down Syndrome become less hypothetical, they are more likely to opt for abortion. 23%-33% of prospective parents, 46%-86% of pregnant women at increased risk for having a child with Down Syndrome, and 89%-97% of women who received a positive diagnosis of Down Syndrome during the prenatal period said they would abort a fetus with Down Syndrome. These findings are similar to our thought experiment in that those whose actions are less likely to be carried out are also more generous.

4. EXPERIMENTAL RESULTS

.

4.1. *Experiment 1*

Figure 4 reports that, on average, participants donated 25% when π was high and 38% when π was low. This result is consistent with mixed consequentialist-deontological motives.

FIGURE 4.— Donation (κ pooled)

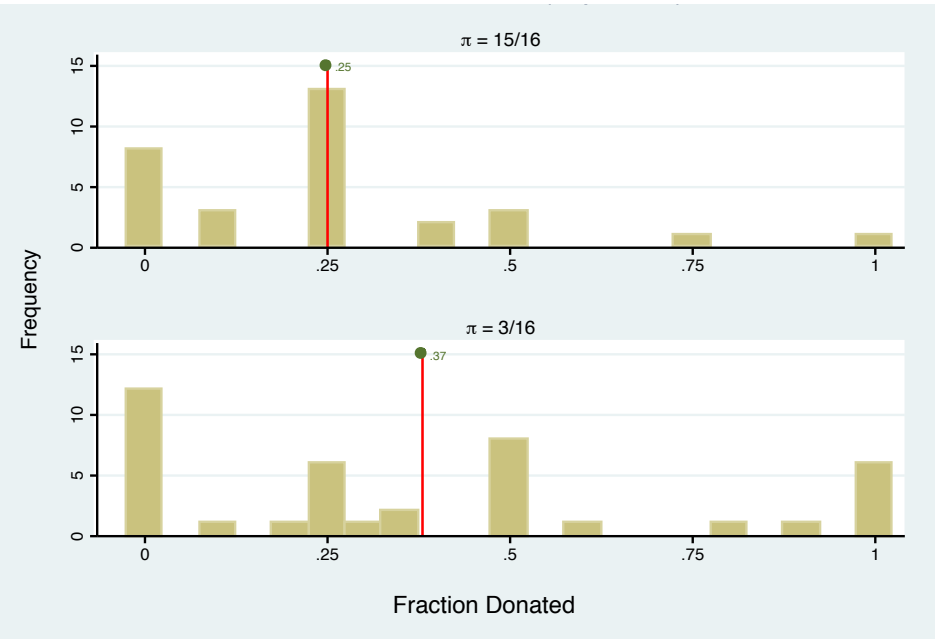


Figure 5 shows that the roughly 50% increase in donations is observed in both $\kappa = 0$ and $\kappa = Max$ treatments, which rejects an ex ante fairness explanation for the results in Figure 4.

FIGURE 5.— Donation (by κ)

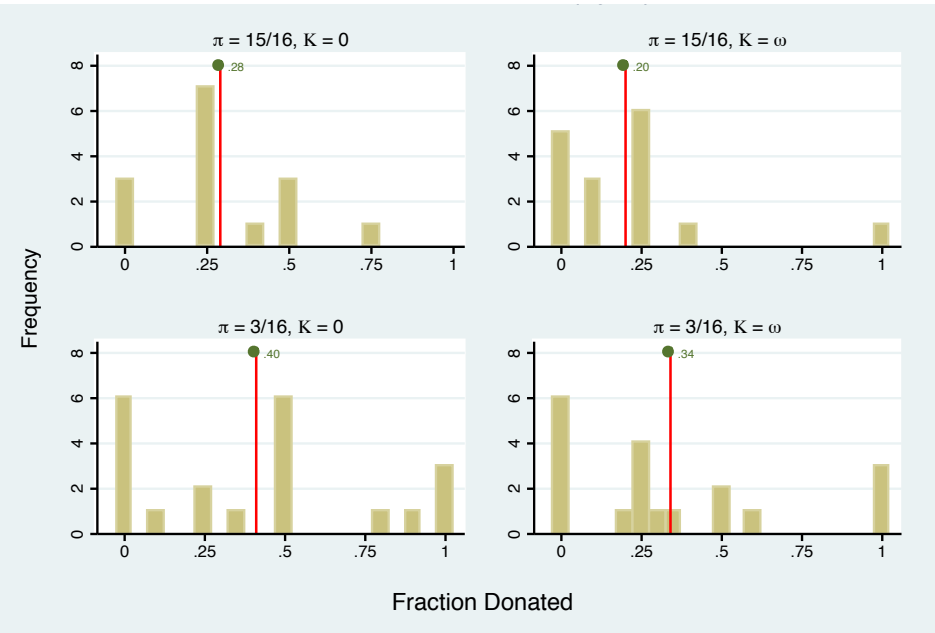


Table 1 reports regression results indicating that the change in donations is significant at the 10% level with or without κ fixed effects (Columns 1 and 2). Increasing the likelihood of implementation from 0% to 100% reduces the donation by 16% points. The remaining columns indicate that subjects are neither targeting expected income nor expected giving. Increasing the likelihood of implementation from 0 to 1 reduces the expected income of the donee by 26% and increases the expected giving of the donor by 22%.

TABLE I

POOLED RESULTS (ZURICH SHREDDING LAB)

	Ordinary Least Squares					
	(1)	(2)	(3)	(4)	(5)	(6)
	d^*		Expected Income E(x_2)		Expected Giving (πd^*)	
Mean dep. var.	0.30		0.39		0.12	
% Consequential (π)	-0.176*	-0.159*	-0.259**	-0.278***	0.212***	0.219***
	(0.0978)	(0.0855)	(0.108)	(0.0802)	(0.0484)	(0.0452)
K Fixed Effects	N	Y	N	Y	N	Y
Observations	71	71	71	71	71	71
R-squared	0.045	0.292	0.077	0.506	0.218	0.339

Notes: Standard errors in parentheses. Raw data shown in Figures 1 and 2. * p < 0.10, ** p < 0.05, *** p < 0.01

Figure 6 graphically examines the ex ante fairness explanation. It shows that as π changes, expected income of the recipient is not fixed; it increases when κ is high and decreases when κ is low.

FIGURE 6.— Expected Income E(x_2) (by κ)

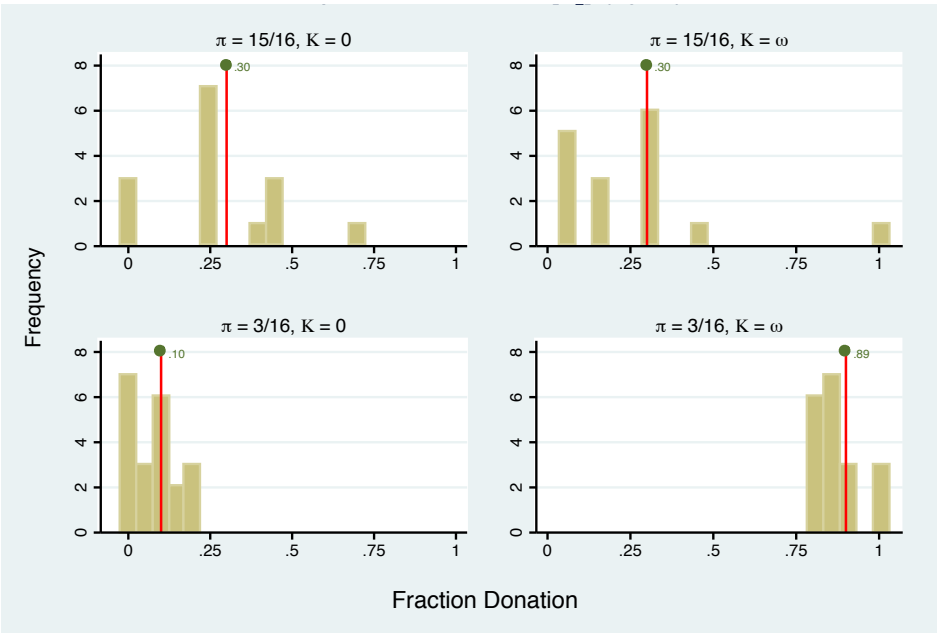
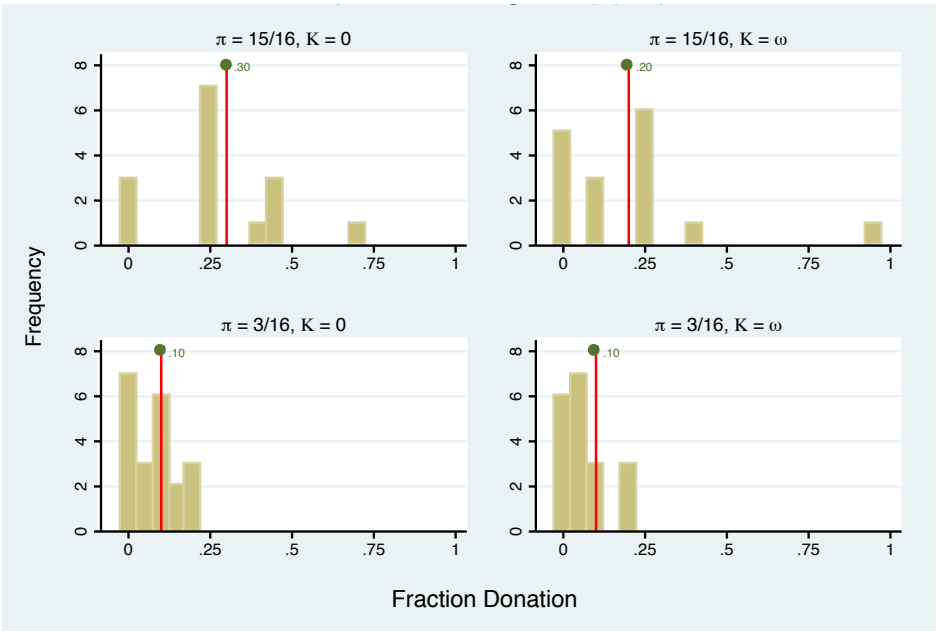


Figure 7 shows that as π changes, expected giving by the decision-maker is also not fixed. This is worth noting as one theory about how decisions are incorporated into self-image or identity involves the expected action of an individual.

FIGURE 7.— Expected Giving (πd^*) (by κ)



Our results indicate that for both κ, expected giving drops by half as π goes from high to low. The statistical significance (1% level) of these results are displayed in Table 1. Wilcoxin-Mann-Whitney

tests show that the distribution of donations as π increases is not significantly different (Table 2) at the 10% level.

TABLE II
NON-PARAMETRIC TESTS (ZURICH SHREDDING LAB)

	Wilcoxon-Mann-Whitney 2-sided test (p-values)
Thresholds	Pooled
$\pi = 3/16$ vs. $\pi = 15/16$	0.16
$K = 0$ vs. $K = \text{Max}$	0.16

4.2. *Experiment 2*

This section reports the results from a larger sample and explores heterogeneity in $\frac{\delta d}{\delta \pi}$ as it correlates with political attitudes, behavior, and vignette responses. Recall that the shredding experiment only reveals the location of one's duty for mixed consequentialist-deontological individuals. When duty to donate is high, increasing the probability of implementation reduces the donation decision ($\frac{\delta d}{\delta \pi} < 0$). When the duty to donate is low (for example, if the duty to self or family is greater than the duty to charitable organizations), then increasing the probability of implementation increases the donation decision ($\frac{\delta d}{\delta \pi} > 0$).

Figure 8 shows that $\frac{\delta d}{\delta \pi} > 0$ among younger individuals (below the median age of 24), non-volunteers,[17] individuals whose political choices are right-wing,[18] and people who responded with errors on the cognitive reflection test (CRT). Individuals who are politically conservative or who do not regularly volunteer may not view charities in a positive light (Greene 2014). Right-wing individuals may feel it is a duty for individuals to be self-sufficient and not rely on charity. Doctors Without Borders may not be well-known among the youngest generation.[19] In these cases, when the probability of implementation is high, donations may be motivated in part from social pressure or experimenter observation. Their ideal decision motivated by duty would be to donate less.

In contrast, older individuals, volunteers, left-wing individuals, and people who answered the CRT correctly tended to increase their donation with a lower probability of implementation ($\frac{\delta d}{\delta \pi} < 0$). The point estimates are between 0.1 to 0.2 in absolute value, indicating that increasing the likelihood of implementation from 0% to 100% reduces the donation by 10-20% points. The differences between groups for each of these four splits of the sample are statistically significant in Figure 8. We only asked nine demographic questions: Four of the demographics show heterogeneous treatment effects while the other five (college major, religion, religious attendance, gender, and whether they donated in the last year) did not. Since four of nine tests yield statistical significance, the heterogeneity we document is unlikely to be a statistical artifact.

[17] Volunteers less frequently than once a month.
[18] We code Right-Wing as CDU/CSU, FDP, AfD, and Other and Left-Wing as SPD, Greens, Pirates, and Lefts.
[19] Scoring higher on standardized tests like the CRT may be correlated with being a left-leaning liberal.

FIGURE 8.— Heterogenous Treatment

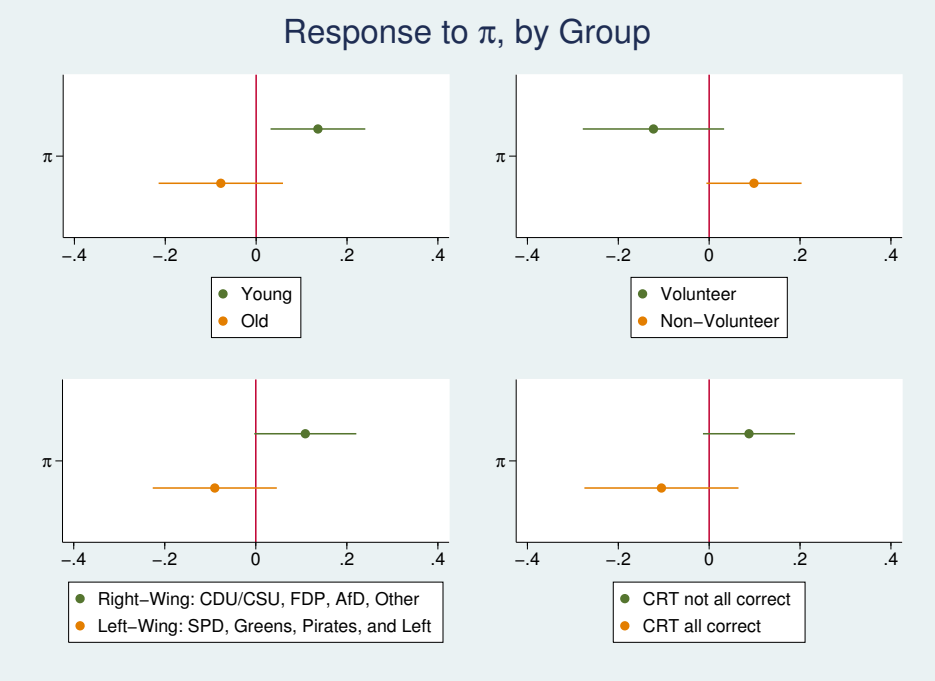
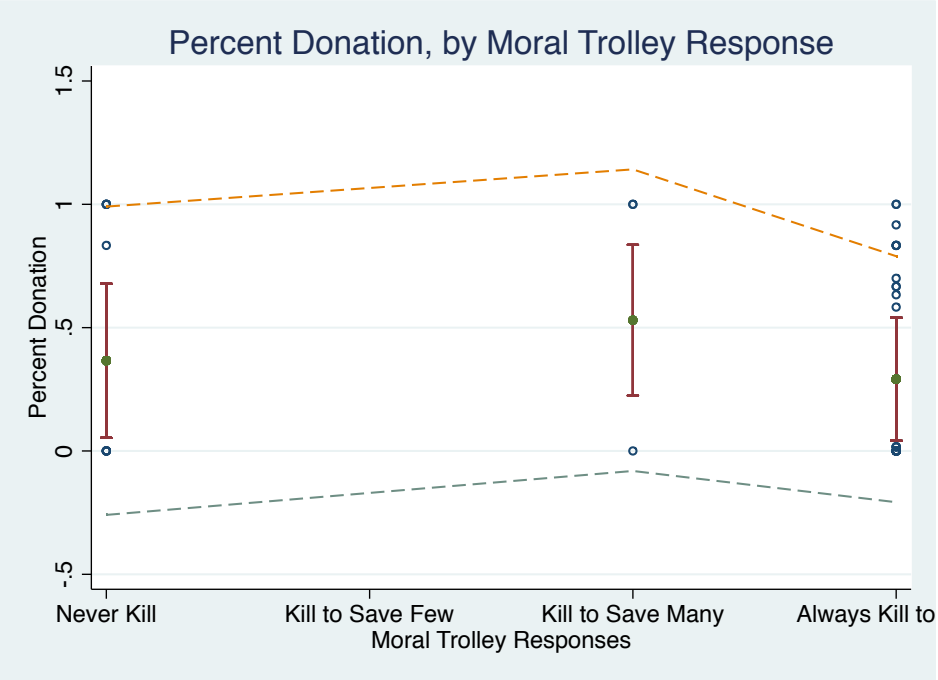


Figure 9 shows that there is no strong pattern of correlation between whether individuals were more likely to respond to the probability of implementation and whether individuals were more likely to change their response on the moral trolley vignette. One reason for this may be sample size, as only 11 individuals whose envelopes were opened reported switching their moral trolley response when the consequences changed. Alternatively, the moral trolley vignette and revealed preference experiment may measure different dimensions of deontological commitments. Indeed, the moral trolley response (duty to not kill) does not appear to be a strong predictor of donations.

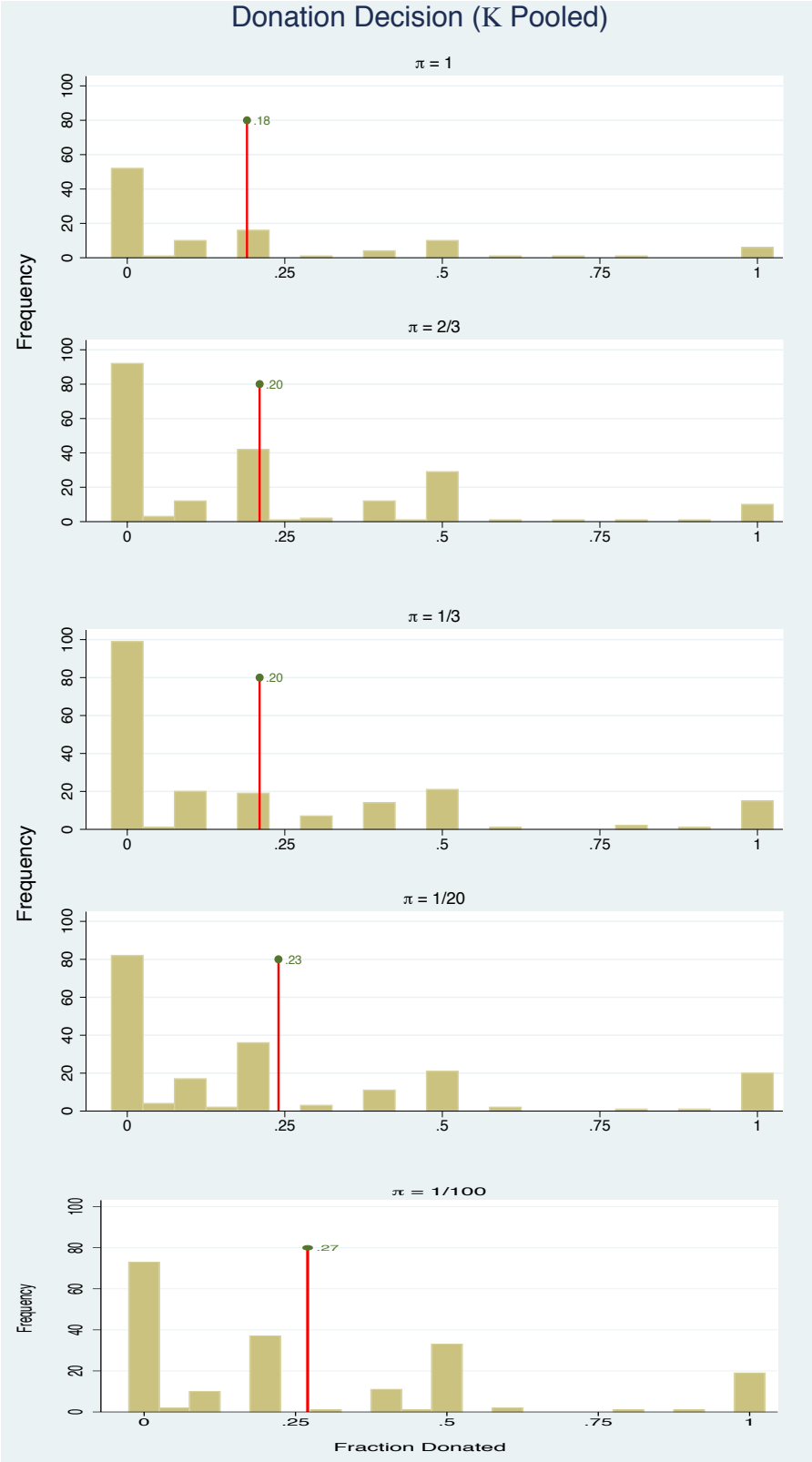
FIGURE 9.—



4.3. *Experiment 3*

Figure 10 reports that in the MTurk sample, the lower the probability that the decision is consequential, the more generous is the decision-maker and the increase in generosity is monotonic with the decrease in probability.

FIGURE 10.— Donation (AMT)



Red: Mean

Donations increased from 18% (when $\pi = 1$) to 27% (when $\pi = 0.01$). Table 3 reports that the effect of π is significant at the 5% level. Columns 3 and 4 examine expected income and Columns 5 and 6 examine expected giving. As in Experiment 1, these quantities are not fixed, suggesting that participants are neither ex ante consequentialists nor targeting expected action. Increasing the likelihood of implementation from 0 to 1 reduces the expected income of the donee by 22% and increases the expected giving of the donor by 20%. Calculating expected income of the donee, we observe strong rejection of ex ante consequentialism: the visual plot and regressions show that the expected donation changes unambiguously with π. To make calculations on expected donations when κ is unknown, we use data on perceived donation.

TABLE III

Pooled Results (AMT)

	Ordinary Least Squares					
	(1)	(2)	(3)	(4)	(5)	(6)
	d^*		Expected Income E(x_2)		Expected Giving (πd^*)	
Mean dep. var.	0.23		0.34		0.07	
% Consequential (π)	-0.0725**	-0.0684*	-0.224***	-0.219***	0.194***	0.213***
	(0.0288)	(0.0390)	(0.0334)	(0.0299)	(0.0132)	(0.0181)
K Fixed Effects	N	Y	N	Y	N	Y
Controls	N	Y	N	Y	N	Y
Observations	902	900	902	900	902	900
R-squared	0.007	0.059	0.048	0.604	0.194	0.214

Notes: Standard errors in parentheses. Raw data shown in Figure 3. Controls include indicator variables for gender, American, Indian, Christian, Atheist, aged 25 or younger, and aged 26-35 as well as continuous measures for religious attendance and accuracy in the lock-in data entry task. * p < 0.10, ** p < 0.05, *** p < 0.01

Table 4 examines each κ treatment arm separately, and we find a quantitatively similar 5.3% to 7.8% decrease as π goes from 0 to 1. Reassuringly, the effects are not significantly different across treatment arms. Other significant predictors of donations are being Indian (who donate 8.4% less than others) and being under 25 (who donate 5.6% less than others).

TABLE IV

DISAGGREGATED RESULTS (AMT)

	Ordinary Least Squares							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Decision (d)		Decision (d)		Decision (d)		Decision (d)	
	K = Unknown		K = 10¢		K = 0¢		K = 50¢	
Mean dep. var.	0.26		0.22		0.20		0.22	
% Consequential (π)	-0.0778	-0.0654	-0.0525	-0.0321	-0.0711	-0.0708	-0.0644	-0.0675
	(0.0523)	(0.0523)	(0.0526)	(0.0536)	(0.0464)	(0.0466)	(0.0462)	(0.0456)
Male		-0.0909**		-0.0474		0.0108		0.0178
		(0.0399)		(0.0430)		(0.0395)		(0.0362)
American		0.0241		-0.0539		0.0838		0.117*
		(0.0524)		(0.0539)		(0.0664)		(0.0598)
Indian		-0.0672		-0.0785		-0.0673		-0.0626
		(0.0566)		(0.0560)		(0.0630)		(0.0590)
Christian		-0.0295		0.0584		-0.0215		-0.000293
		(0.0483)		(0.0503)		(0.0494)		(0.0479)
Atheist		-0.0188		0.00480		0.0113		-0.0927
		(0.0644)		(0.0649)		(0.0802)		(0.0725)
Religious Services Attendance		-0.00614		0.000508		0.00367		-0.00546
		(0.0145)		(0.0156)		(0.0137)		(0.0137)
Ages 25 or Under		-0.0207		-0.122**		-0.0109		-0.113**
		(0.0518)		(0.0570)		(0.0493)		(0.0474)
Ages 26-35		0.00271		-0.110*		-0.00105		-0.111**
		(0.0548)		(0.0593)		(0.0493)		(0.0480)
Own Errors		-0.000192		-0.000186		0.000220		-0.000148
		(0.000193)		(0.000163)		(0.000194)		(0.000143)
Observations	260	260	218	218	256	255	271	270
R-squared	0.009	0.069	0.005	0.081	0.009	0.052	0.007	0.097

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

We next examine whether the distribution of donation decisions is significantly affected by π. Table 5 shows that along most thresholds for π, the distribution of donations as π increases is significantly different.

TABLE V

NON-PARAMETRIC TESTS (AMT)

	Wilcoxon-Mann-Whitney 2-sided test (p-values)		
	(1)	(2)	(3)
Thresholds	K Unknown or 10¢	K = 0¢ or 50¢	K Pooled
$\pi = 1$ vs. $\pi \leq 0.67$	0.91	0.05	0.11
$\pi \geq 0.67$ vs. $\pi \leq 0.33$	0.07	1.00	0.20
$\pi \geq 0.33$ vs. $\pi \leq 0.05$	0.05	0.10	0.01
$\pi \geq 0.05$ vs. $\pi = 0.01$	0.15	0.02	0.01

	π Pooled
$K \geq$ 10¢ vs. K = 0¢	0.40
K = 50¢ vs. $K \leq$ 10¢	0.11

To interpret Table 5, 0.05 in Column 1 means that we reject with 95% confidence the hypothesis that the distribution of decisions for people treated with $\pi = 1, 0.67, 0.33$ is the same as the distribution of decisions for people treated with $\pi = 0.05, 0.01$. Table 5 also reports that the distribution of donations does not significantly vary by κ. Qualitatively similar results are found in the shredding experiment; differences by π are more significant than differences by κ. Means are also

not significantly different by κ in either experiment. Note that FOSD makes no strong predictions about $\frac{\partial d^*}{\partial \kappa}$.

4.4. *Heterogeneity*

Examining the raw data suggests that there is substantial heterogeneity and that there may be many people who do not respond to treatment, i.e. they always donate 0%, 50%, or 100%. One interpretation of our results could be that there are sizeable fractions of people who are pure consequentialist or pure deontological and a large fraction of people who have hybrid motivations. We now examine heterogeneity as it correlates with observable demographics.

Table 6 shows that along all demographic groups, $\frac{\delta d}{\delta \pi} < 0$. Americans, Christians, Atheists, and those who are less likely to attend religious services are particularly likely to have steeper $\frac{\delta d}{\delta \pi}$.

TABLE VI

WHO RESPONDS TO π? (AMT)

	Ordinary Least Squares									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Decision (d)									
Mean dep. var.	0.23									
% Consequential (π)	-0.100**	-0.0493	-0.124**	-0.0500	-0.0522	-0.0774	-0.0618	-0.0548	-0.0839**	-0.0190
	(0.0494)	(0.0429)	(0.0506)	(0.0436)	(0.0403)	(0.0616)	(0.0467)	(0.0443)	(0.0407)	(0.126)
π * Male	0.0612									0.0490
	(0.0577)									(0.0611)
π * American		-0.0675								0.0370
		(0.0627)								(0.0911)
π * Indian			0.0990*							0.0426
			(0.0574)							(0.0963)
π * Christian				-0.0599						-0.0658
				(0.0632)						(0.0783)
π * Atheist					-0.133					-0.145
					(0.0837)					(0.108)
π * Religious Services Attendance						0.00394				-0.00739
						(0.0210)				(0.0224)
π * Ages 25 or Under							-0.0149			-0.0815
							(0.0576)			(0.0787)
π * Ages 26-35								-0.0386		-0.0878
								(0.0597)		(0.0808)
π * Own Errors									0.000402	0.000319
									(0.000299)	(0.000307)
K Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	900	900	900	900	900	900	900	900	900	900
R-squared	0.061	0.061	0.063	0.060	0.062	0.059	0.059	0.060	0.061	0.068

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Column 3 displays a significant coefficient on the interaction with being Indian that is positive, though this significant interaction may be an artifact of multiple hypothesis testing. Summing this interaction term with the level effect indicates that Indians (40% of the sample) are more pure consequentialist or deontological than others.

Even when all covariates' interactions are included, Atheists appear to be the most mixed consequentialist-deontological in their motivations. While we did not ask about political orientation in Experiment 3, this is consistent with Experiment 2, which found left-wing individuals, who tend to be less religious (Chen and Lind 2014), to increase their donations more when the probability of implementation was small. Moreover, the steeper $\frac{\delta d}{\delta \pi}$ among wealthier individuals (i.e. Americans) is also consistent with comparing the results of Experiments 1 and 2, as the subjects in Zurich also displayed steeper $\frac{\delta d}{\delta \pi}$ than subjects from the relatively poorer Hamburg and Magdeburg (which is

part of the former East Germany).

4.5. *Cognition Costs and Diffusion of Responsibility*

Under the cognitive cost model, individuals spend less time thinking and possibly use altruistic heuristics when their decision is less likely to be implemented. However, Figure 11 shows that individuals in Zurich spend roughly the same time thinking about their decision regardless of the implementation probability. Moreover, subjects do not donate less when they spend more time on their decision.

FIGURE 11.— Time Spent (Zurich Lab)

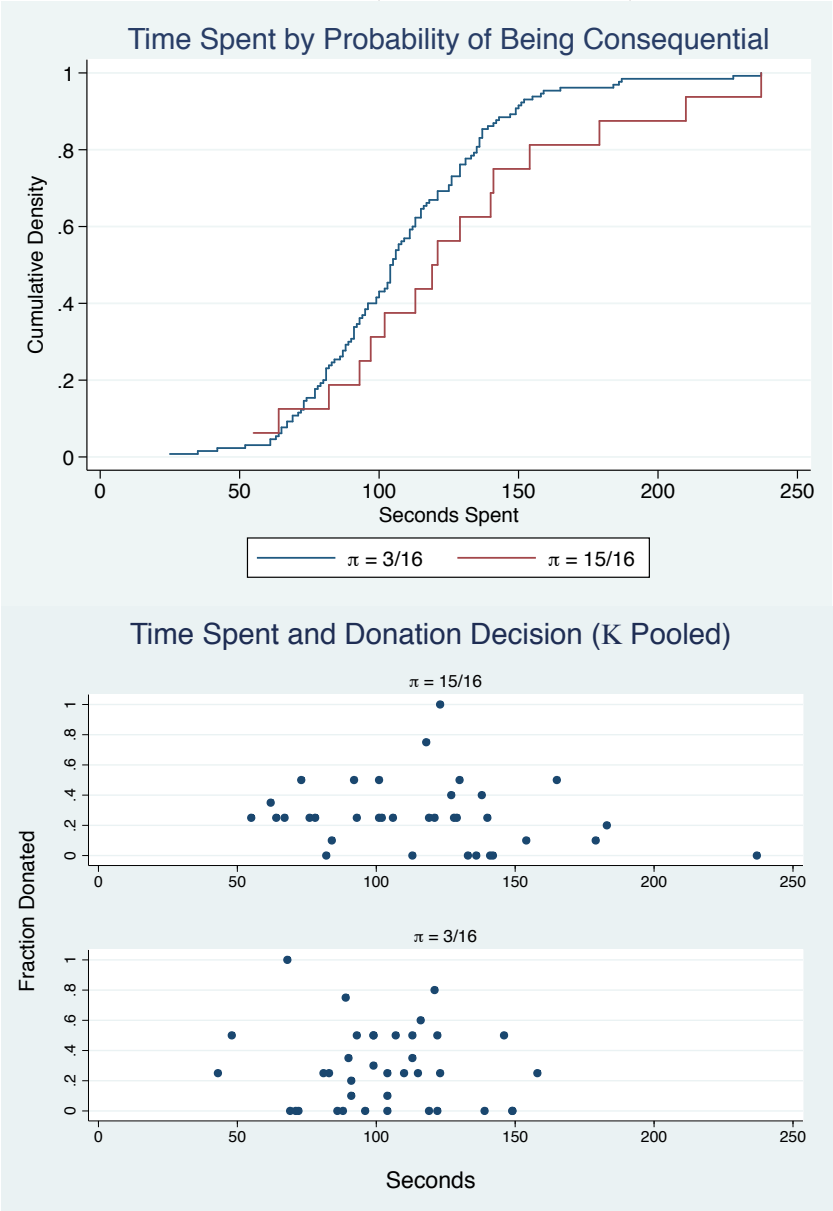


FIGURE 12.— Time Spent (AMT Experiment)

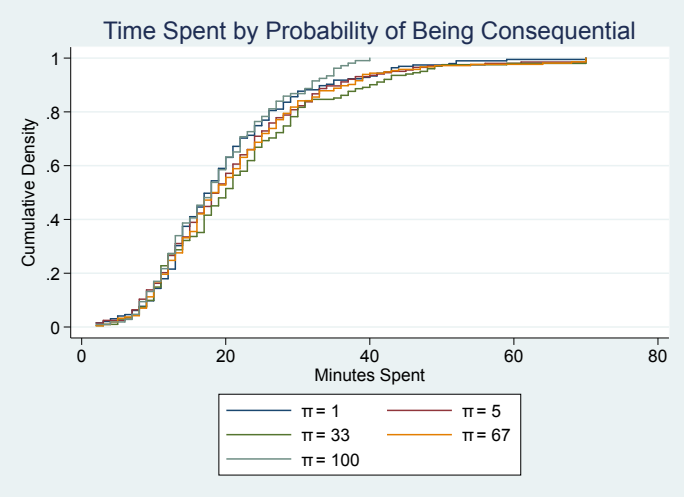


Table 7: Time for Completion of Survey (AMT Experiment)

Sample	All Subjects	Above Median Mixed-Consequentialist		Below Median Mixed-Consequentialist	
	(1)	(2)	(3)*	(4)	(5)*
Mean dep. var.			20.8		
% Consequential (π)	0.0123	0.0176	0.0452	0.163***	0.118*
	(0.0162)	(0.0547)	(0.0574)	(0.0548)	(0.0635)
π^2		-0.000482	-0.000452	-0.00167***	-0.00122*
		(0.000573)	(0.000602)	(0.000581)	(0.000674)
Above Median Mixed-Consequentialist	0.755				
	(1.119)				
π * Above Median Mixed-Consequentialist	-0.0386*				
	(0.0227)				
Observations	900	449	449	451	451
R-squared	0.004	0.008		0.019	

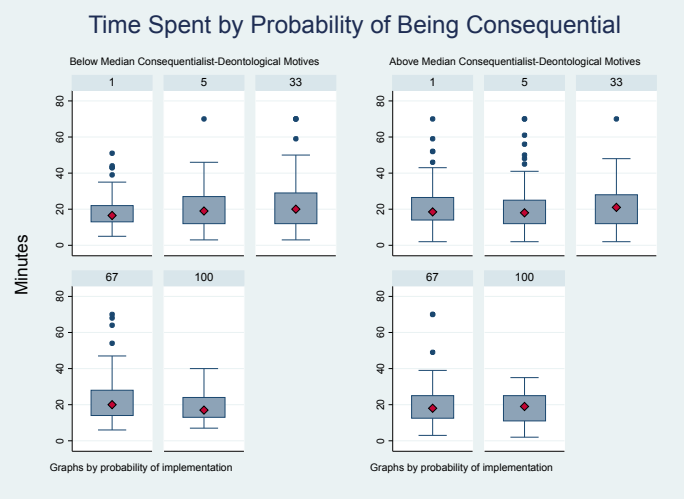
Notes: Standard errors in parentheses. Mixed-Consequentialist aggregates for each subject their demographic characteristics' contribution to the effect of π on the Donation decision. Regressions are weighted by the standard deviation of the first regression to account for uncertainty in the calculation of mixed-consequentialist score. Columns 3 and 5 employ median regressions. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 12 shows the analogous results on MTurk: Time spent is only affected (and reduced) by $\pi = 1$. This result would appear inconsistent with cognition costs. Donations were again not associated with time spent, but would be negatively associated under a theory that cognition costs explain increased generosity when the implementation probability is low.

One prediction from the cognition cost model is that those whose behavior is most elastic to π should resort to heuristics more when the probability of being consequential is low. However, Table 7 shows that at low π, those with below-median $\frac{\delta d}{\delta \pi}$ spend less time than those with above-median $\frac{\delta d}{\delta \pi}$. In addition, Figure 13 shows that those with high $\frac{\delta d}{\delta \pi}$ do not vary time spent as π changes.

Patterns in our data also reject alternative explanations such as diffiusion of responsibility: If someone feels less responsible for the outcome, they may choose to be more selfish, under the argument that the lottery chose the final outcome for the recipient. Decision-makers should become

FIGURE 13.— Time Spent by $\frac{\delta d}{\delta \pi}$ (AMT Experiment)



Red Diamond: Median

more generous with a high probability of implementation. Another alternative explanation is loss of control (Fehr et al. 2013). If individuals value authority, they may compensate themselves for the loss of control, which would also predict decisions to become more generous with a high probability of implementation. However, we find the opposite.

4.6. *Trading Off Consequentialist and Deontological Motivations*

If we make functional form assumptions about consequentialist and deontological motivations, we can obtain estimates about how individuals trade off between consequentialist and deontological motivations. Our goal is to write the first-order condition for individuals' utility, treat the data as if they are the outcome of utility maximization, and then estimate the parameters that achieve the maximum likelihood for the observed data. In particular, the first-order conditions provide moment conditions that we try to fit. Since we are interested in the first-order condition with respect to individuals' decisions, we can focus on the decision-dependent portion of expected utility.

We consider two cases.

i) Consequentialist motivations are homo oeconomicus.

$u(x_{DM},, x_2, d) = \lambda(x_1) + (-(\delta - d)^2) = \lambda(1-d) + (-(\delta - d)^2)$

The deontological portion uses bliss point preferences. This formulation is similar to Cappelen et al. (2007, 2013), meaning that subjects view their duty as $d = \delta$ rather than $d \geq \delta$.

ii) Consequentialist motivations are Fehr-Schmidt.

$u(x_{DM},, x_2, d) = \lambda(x_1 - \alpha max\{x_2 - x_1, 0\} - \beta max\{x_1 - x_2, 0\}) + (-(\delta - d)^2)$.

In both cases, we would like to estimate the bliss point, δ, and the relative weight individuals place on the consequentialist motivations, λ. Decisions and outcomes are in percentages as donations in different experiments exhibit modes at certain fractions, like 50% or 25%. We believe subjects' duties are enumerated in percent terms. Consequentialist motivations can easily be enumerated in

cents and λ would then represent the trade-off between cents and fractions.[20]

4.6.1. *Homo Oeconomicus and Deontological Motivations*

The first-order condition is: $0 = \pi\lambda(-1)+2(\delta-d)$. This results in a linear regression, $-\frac{\lambda}{2}\pi+\delta = d^*$. Note that we can interpret the constant term of the linear regression as the bliss point. This is intuitive since the constant term represents the decision when $\pi = 0$. We can precisely estimate this term as 25%. Our estimate of -0.073 from Table 2 implies that $\lambda = 0.14$. This small weight is intuitive since the data reveals that many people donate more than the bliss point of 25%.

4.6.2. *Fehr-Schmidt and Deontological Motivations*

In principle, we would like to separately estimate the bliss point, δ, the weight individuals place on the consequentialist motivations, λ, and the inequality parameters, α and β. We plug in d for x_2, x_1: $\pi\lambda(1 - d - \alpha max\{d - (1 - d), 0\} - \beta max\{(1 - d) - d, 0\}) + (-(\delta - d)^2)$.

We can rewrite this as: $\pi\lambda(1 - d - \alpha max\{2d - 1, 0\} - \beta max\{1 - 2d, 0\}) + (-(\delta - d)^2)$. This expression is quadratic in d, so the first-order condition, and hence moment conditions, will be linear in d. Thus, we will be estimating a linear regression to back out our parameters of interest. To see this, first observe that the decision-dependent portion of expected utility if $\frac{1}{2} > d$, is: $\pi\lambda(1 - d - \beta(1 - 2d)) + (-(\delta - d)^2)$, else $\pi\lambda(1 - d - \alpha(2d - 1)) + (-(\delta - d)^2)$.

The individual's first-order condition over their choice d is then given by the following expression. If $\frac{1}{2} > d$, then: $0 = \pi\lambda(2\beta - 1) + 2(\delta - d)$, else $0 = \pi\lambda(-2\alpha - 1) + 2(\delta - d)$.

Thus, our linear regression is: If $\frac{1}{2} > d$, then $\pi\frac{\lambda(2\beta-1)}{2} + \delta = d^*$, else $\pi\frac{\lambda(-2\alpha-1)}{2} + \delta = d^*$. This expression motivates our GMM condition:

$E\left[\pi\left(1[\frac{1}{2} > d]\left[d - \pi\frac{\lambda(2\beta-1)}{2} - \delta\right] + 1[\frac{1}{2} \leq d]\left[d - \pi\frac{\lambda(-2\alpha-1)}{2} - \delta\right]\right)\right] = 0.$

Equivalently, we can run a linear regression of d on $1[\frac{1}{2} > d]\pi$ and $1[\frac{1}{2} \leq d]\pi$. However, the ordinary least squares version of this regression is somewhat problematic because the decision appears on both the left-side of the equation as outcome and on the right-side in the indicator function, which would drive a spurious correlation on β_2 were we to estimate $d_i = \beta_0 + \beta_1\pi_i + \beta_2 1[\frac{1}{2} \leq d_i]\pi_i + \varepsilon_i$. We thus need to instrument for $1[\frac{1}{2} \leq d_i]$ that is not directly correlated with d_i.

Estimates using two different instruments, being Indian or being under 25, results in similar point estimates (Table 8). The bliss point is to donate 25% of endowment. The first coefficient indicates that while $d < 50\%$, donation increases as π decreases. However, once $d > 50\%$, donation decreases as π decreases. This is intuitive. Since the bliss point for duty is below 50%, then for people to meet their duty as π falls, they should be moving towards 25% donation, which is less than 50%.

Our results suggest that $\frac{\lambda(2\beta-1)}{2} = -0.36$ and $\frac{\lambda(-2\alpha-1)}{2} = 1.16$. With two equations and three unknowns, we cannot identify our parameters. However, we can choose values for β and α in the range of values in Fehr and Schmidt (1999). But, if individuals are inequality averse and are more averse to adverse inequality, we know that $\alpha > \beta > 0$; examining $\frac{\lambda(-2\alpha-1)}{2} = 1.16$ implies $\lambda < 0$. Since $\lambda = 0$ as a boundary condition, these calculations would suggest that, in our experiment,

[20] $u(x_{DM,}, x_2, d) = \lambda\omega(x_1) + (-(\delta - d)^2) = \lambda\omega(1 - d) + (-(\delta - d)^2)$

Table 8: Trading Off Consequentialist-Deontological Motivations (AMT Experiment)

	OLS	IV	IV
	(1)	(2)	(3)
	Decision (d)		
Mean dep. var.	0.23		
% Consequential (π)	-0.239***	-0.363***	-0.368***
	(0.0249)	(0.0548)	(0.139)
π * 1($d \geq w/2$)	0.870***	1.516***	1.542**
	(0.0412)	(0.250)	(0.714)
Constant (Duty Bliss Point)	0.251***	0.249***	0.249***
	(0.0116)	(0.0131)	(0.0134)
IV	N	π, Indian	π, Age $\leq$ 25
Observations	902	902	902
R-squared	0.336	0.155	0.140

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

behavior that may appear as consequentialist Fehr-Schmidt preferences may be largely explained by deontological motivations.

5. IMPLICATIONS FOR EXPERIMENTAL METHODS

If deontological motivations exist, our model and results present a critique of the random lottery incentive and the strategy method frequently used in experimental economics to collect additional data. These methods elicit many decisions from decision-makers, but only one decision will be implemented. If decision-makers view this decision deontologically, then, decision-makers may report more moral decisions than would be the case than if the decision was definitely implemented, and in experiments that randomize treatment conditions, differences across treatment conditions may be magnified. A similar critique bears on policymakers' use of contingent valuation and psychologists' use of vignettes. A formal argument on the implications of deontological motivations for experimental methods is made in Chen and Schonger (2014).

Several prominent papers already interrogate the random lottery incentive. A seminal paper supports its use by showing that utility must be approximately linear (Rabin 2000); thus, if expected utility also holds, then when only one decision is payoff-relevant, risk aversion should not affect decision-making in these games. Our results show that even if utility is approximately linear, changes in the probability of decisions being payoff-relevant should affect decision-making when the decision has a moral element. Holt (1986), a well-known theoretical critique of the random lottery incentive, shows that if subjects understand the whole experiment as a single game and violate the independence axiom, considering each experiment by itself does not give subjects' true preferences. This critique already applies for decision-problems where there is no other player and no moral dimension whatsoever, as in choices among lotteries. Starmer and Sugden (1991) experimentally test whether this potential problem identified by theory is a problem in practice, and conclude that it is not, thus our critique may be limited to experiments on social/moral preferences. Contingent valuation studies that query individuals' hypothetical preferences have already been criticized for non-formal reasons (Diamond and Hausman 1994; List and Gallet 2001; List 2001).

6. CONCLUSION

In recent decades, behavioral economics has shown that individuals make decisions not solely based on self-interest–that is, only considering consequences on oneself–but also based on the consequences for others. This paper provides clean experimental evidence that focusing solely on consequences is too narrow; rather, individuals also seem to care about decisions in-and-of-themselves, independently, of consequences. The traditional approach to measuring consequentialist vs. deontological preferences is through vignette studies (Greene et al. 2001; Chen 2012). We derive predictions from a simple, but general, model of social preferences to infer the presence of consequentialist or deontological preferences (or both) in an actual experiment involving costly actions. In several experiments, we investigate whether individuals care about actions per se rather than about the consequences of actions. We begin with a formal investigation of whether individual pro-social behavior varies with the likelihood that their pro-social decision will actually be implemented. We show that any change is inconsistent with standard behavioral preferences. The intuition is that preferences that depend only on outcomes, whether directly *or indirectly*, would predict that decisions (altruism, truth-telling, promise-keeping) are constant in the probability because varying the probability equally affects the benefits and costs of some action.

We find that individuals share more with charitable organizations when the likelihood that their sharing decision will actually be implemented is lower. Our results suggest that people care about decisions in-and-of-themselves. Decisions are remembered and relevant even when they are inconsequential–in the strong sense that they not only do not affect payoffs, but moreover when other agents never learn about them. On a fundamental level, our results suggest that deontological motivations can explain some of the patterns previously attributed to consequences. Observations linking social preferences with outcomes, in some cases, may be due to individuals simply being hardwired to display those preferences even with no possibility of punishment or reward. Empathic concern, duty-driven, deontological decision-making may occur, regardless of the consequences for the potential beneficiaries. Behaviorally, such effects may be more prevalent than previously thought. The existence of mixed consequentialist-deontological individuals would be consistent with evolutionary models that predict convex combinations of homo oeconomicus and homo kantiensis preferences will be evolutionarily stable.

Our paper can also be distinguished from the prior psychology and philosophy literature by providing a revealed preference method to detect deontological motivations that is typically interpreted from self-reported data or hypothetical vignettes. We do not find the vignette-approach to align with revealed preferences, though further research should investigate whether this is due to sample size or due to the measures capturing different dimensions of deontological motivations. Our paper can also be distinguished from the prior economics literature by measuring a narrower, internal audience, motivation for observed social preferences rather than social audience motivation, and from the prior experimental literature by offering a method to rule out a strong form of experimenter demand (subjects inferring that the experimenter expects a certain outcome) and a weak form of experimenter demand (subjects are motivated by the knowledge their decision is consequential to

science). As far as we are aware, neither form of experimenter demand has been addressed.

Future research should examine whether deontological motivations differ between individuals and how deontological motivations come about. Any behavioral change in the experiment is inconsistent with standard behavioral preferences that depend directly or indirectly on outcomes only. With functional form assumptions, the direction of change reveals the location of one's greatest duty for a mixed consequentialist-deontological decision maker. We see some heterogeneity in the location of duty among subjects in large samples and comparing across experiments. In recent years, economists have begun trying to understand repugnance as a rejection of optimal institutional arrangements (Roth 2007; Mankiw and Weinzierl 2010) and, in the lab, have begun documenting heterogeneity in behaviors that resemble deontological commitments, such as preferences for procedural fairness (Gibson et al. 2013; Brock et al. 2013). Some suggest that war is about a conflict of "sacred values" (Bowles and Polania-Reyes 2012; Chen 2006, 2010; Berdejó and Chen 2013) and that legal compliance is driven in part by perceived legitimacy of law (Tyler and Huo 2002; Chen 2013). The U.S. Department of Defense has begun a Minerva Initiative to understand, and perhaps change, sacred values.[21] An interesting subject of future research is welfare economics or policy responses that incorporate deontological motivations (Chen et al. 2011; Chen 2004; Chen et al. 2014) or the causal effects of deontological motivations.[22]

In law, we are sometimes interested in the motivations (mental state) of the litigant, e.g., in copyright disputes, judges may care about behavior motivated by a creator's moral rights, in equity law, judges may care about opportunistic behavior, and most famously in criminal law when a distinction is made between *mens rea* (intention) and *actus reus* (act). In managerial settings, some stress the importance of how to screen or select for the presence of deontological motivations in business leaders, politicians, or judges (Besley 2005; Chen et al. 2015). For philosophers who argue that human dignity derives from the possibility of deontological decision-making and for theorists who postulate the existence of deontological motivations (Bénabou and Tirole 2011; Alger and Weibull 2012), linking their theoretical predictions about the positive prevalence of deontological motivations to results about the actual prevalence of deontological motivations from studies using our revealed preference method for detecting them seems to be a promising research program.

[21] There is some field evidence consistent with the malleability of sacred values (Chen and Yeh 2014; Chen and Givati 2014; Chen 2014; Chen and Yeh 2013b).

[22] For example, random assignment of judges with different deontological commitments would be one way to estimate the causal effects of deontological commitments (Chen and Sethi 2011; Chen and Yeh 2013a).

REFERENCES

Alexander, L. and M. Moore (2012): *Stanford Encyclopedia of Philosophy*.

Alger, I. and J. Weibull (2012): "Homo Moralis-Preference Evolution Under Incomplete Information and Assortative Matching," TSE Working Papers 12-281, Toulouse School of Economics (TSE).

Andreoni, J. (1990): "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *The Economic Journal*, 100, 464–477.

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

Arrow, K. J. (2012): *Social Choice and Individual Values*, Cowles Foundation Monographs Series, New Haven: Yale university press, 3rd ed., monograph 12.

Batson, C. D., J. G. Batson, J. K. Slingsby, K. L. Harrell, H. M. Peekna, and R. M. Todd (1991): "Empathic Joy and the Empathy-Altruism Hypothesis," *Journal of Personality and Social Psychology*, 61, 413–426.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *The American Economic Review*, 96, 1652–1678.

Bénabou, R. and J. Tirole (2011): "Identity, Morals, and Taboos: Beliefs as Assets," *The Quarterly Journal of Economics*, 126, 805–855.

Bentham, J. (1791): *Panopticon*, London: T. Payne.

Berdejó, C. and D. L. Chen (2013): "Priming Ideology? Electoral Cycles without Electoral Incentives among Elite U.S. Judges," Tech. rep., ETH Zurich, Mimeo.

Bergstrom, T. C., R. J. Garratt, and D. Sheehan-Connor (2009): "One Chance in a Million: Altruism and the Bone Marrow Registry," *The American Economic Review*, 99, 1309–1334.

Besley, T. (2005): "Political selection," *The Journal of Economic Perspectives*, 19, 43–60.

Bowles, S. and S. Polania-Reyes (2012): "Economic Incentives and Social Preferences: Substitutes or Complements?" *Journal of Economic Literature*, 50, 368–425.

Brock, J. M., A. Lange, and E. Y. Ozbay (2013): "Dictating the Risk: Experimental Evidence on Giving in Risky Environments," *The American Economic Review*, 103, 415–437.

Cappelen, A. W., A. D. Hole, E. Ø. Sørensen, and B. Tungodden (2007): "The Pluralism of Fairness Ideals: An Experimental Approach," *American Economic Review*, 97, 818–827.

Cappelen, A. W., J. Konow, E. Ø. Sørensen, and B. Tungodden (2013): "Just luck: An experimental study of risk-taking and fairness," *The American Economic Review*, 103, 1398–1413.

Chen, D. L. (2004): "Gender Violence and the Price of Virginity: Theory and Evidence of Incomplete Marriage Contracts," Working paper, University of Chicago, Mimeo.

——— (2006): "Islamic Resurgence and Social Violence During the Indonesian Financial Crisis," in *Institutions and Norms in Economic Development*, ed. by M. Gradstein and K. A. Konrad, MIT Press, chap. 8, 179–199.

——— (2010): "Club Goods and Group Identity: Evidence from Islamic Resurgence during the Indonesian Financial Crisis," *The Journal of Political Economy*, 118, 300–354.

——— (2012): "Markets and Morality: How Does Competition Affect Moral Judgment," Working paper, Duke University School of Law.

——— (2013): "The Deterrent Effect of the Death Penalty? Evidence from British Commutations During World War I," Working paper, ETH Zurich, Mimeo.

——— (2014): "Can Markets Stimulate Rights? On the Alienability of Legal Claims," Tech. rep.

Chen, D. L. and Y. Givati (2014): "Can Markets Overcome Repugnance? Muslim Trade Reponse to Anti-Muhammad Cartoons," Working paper, ETH Zurich, Mimeo.

Chen, D. L. and J. J. Horton (2014): "The Wages of Pay Cuts," Working paper, ETH Zurich.

Chen, D. L., V. Levonyan, S. E. Reinhart, and G. B. Taksler (2014): "Mandatory Disclosure: Theory and Evidence from Industry-Physician Relationships," Working paper, mimeo.

Chen, D. L., V. Levonyan, and S. Yeh (2011): "Do Policies Affect Preferences? Evidence from Random Variation in Abortion Jurisprudence," Manuscript.

Chen, D. L. and J. T. Lind (2014): "The Political Economy of Beliefs: Why Fiscal and Social Conservatives and Liberals Come Hand-in-Hand," Working paper.

Chen, D. L., M. Michaeli, and D. Spiro (2015): "Preferences for Perfection: Public vs. Private Truths on Judicial Panels," Working paper, ETH Zurich, Mimeo.

Chen, D. L. and M. Schonger (2014): "Invariance of Equilibrium to the Method of Elicitation and Implications for Social Preferences," Working paper, ETH Zurich, Mimeo.

Chen, D. L. and J. Sethi (2011): "Insiders and Outsiders: Does Forbidding Sexual Harassment Exacerbate Gender Inequality?" Working paper, University of Chicago.

Chen, D. L. and S. Yeh (2013a): "Growth Under the Shadow of Expropriation? The Economic Impacts of Eminent Domain," Working paper, Duke University, Mimeo.

——— (2013b): "How Do Rights Revolutions Occur? Free Speech and the First Amendment," Working paper, ETH Zurich, Mimeo, Zurich.

——— (2014): "The Construction of Morals," *Journal of Economic Behavior and Organization*, 104, 84–105.

Choi, H., M. Van Riper, and S. Thoyre (2012): "Decision making following a prenatal diagnosis of Down syndrome: an integrative review," *Journal of Midwifery & Womens Health*, 57, 156–164.

DellaVigna, S., J. A. List, and U. Malmendier (2013): "Voting to Tell Others," Working paper.

Diamond, P. A. and J. A. Hausman (1994): "Contingent Valuation: Is Some Number better than No Number?" *The Journal of Economic Perspectives*, 8, 45–64.

Falk, A. and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54, 293–315.

Feddersen, T., S. Gailmard, and A. Sandroni (2009): "Moral Bias in Large Elections: Theory and Experimental Evidence," *The American Political Science Review*, 103, 175–192.

Fehr, E., H. Herz, and T. Wilkening (2013): "The Lure of Authority: Motivation and Incentive Effects of Power," *The American Economic Review*, 103, 1325–1359.

Fehr, E. and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114, 817–868.

Friedman, M. and L. J. Savage (1948): "The Utility Analysis of Choices Involving Risk," *The Journal of Political Economy*, 56, 279–304.

Gibson, R., C. Tanner, and A. F. Wagner (2013): "Preferences for Truthfulness: Heterogeneity among and within Individuals," *The American Economic Review*, 103, 532–548.

Gneezy, U. (2005): "Deception: The Role of Consequences," *The American Economic Review*, 95, 384–394.

Greene, J. (2014): *Moral tribes: emotion, reason and the gap between us and them*, Atlantic Books Ltd.

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen (2001): "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science*, 293, 2105–2108.

Holt, C. A. (1986): "Preference Reversals and the Independence Axiom," *The American Economic Review*, 76, 508–515.

Kant, I. (1797): "Über ein vermeintes Recht aus Menschenliebe zu lügen," *Berlinische Blätter*, 1, 301–314.

——— (1959): *Foundations of the Metaphysics of Morals*, Prentice Hall, II ed., trans. by L. W. Beck.

Kaplow, L. and S. Shavell (2009): *Fairness versus welfare*, Harvard university press.

Kreps, D. M. (1988): *Notes on the Theory of Choice*, Westview Press Boulder.

Levhari, D., J. Paroush, and B. Peleg (1975): "Efficiency Analysis for Multivariate Distributions," *The Review of Economic Studies*, 42, 87–91.

List, J. A. (2001): "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards," *The American Economic Review*, 91, 1498–1507.

List, J. A. and C. A. Gallet (2001): "What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?" *Environmental and Resource Economics*, 20, 241–254.

Machina, M. J. (1982): ""Expected Utility" Analysis without the Independence Axiom," *Econometrica*, 50, 277–323.

——— (1989): "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty," *Journal of Economic Literature*, 27, 1622–1668.

Mankiw, N. G. and M. Weinzierl (2010): "The Optimal Taxation of Height: A Case Study of Utilitarian Income Redistribution," *American Economic Journal: Economic Policy*, 2, 155–176.

McCabe, K. A., M. L. Rigdon, and V. L. Smith (2003): "Positive reciprocity and intentions in trust games," *Journal of Economic Behavior & Organization*, 52, 267–275.

Mikhail, J. (2007): "Universal Moral Grammar: Theory, Evidence and the Future," *Trends in Cognitive Sciences*, 11, 143 – 152.

Nagel, T. (1970): *The Possibility of Altruism*, Oxford: Clarendon Press.

Nozick, R. (1974): *Anarchy, State, and Utopia*, Harper Torchbooks, Basic Books.

Orne, M. T. (1962): "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist*, 17, 776–783.

Quiggin, J. (1982): "A Theory of Anticipated Utility," *Journal of Economic Behavior & Organization*, 3, 323–343.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *The American Economic Review*, 83, 1281–1302.

——— (2000): "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," *Econometrica*, 68, 1281–1292.

Rand, D. G., J. D. Greene, and M. A. Nowak (2012): "Spontaneous Giving and Calculated Greed," *Nature*, 489, 427–430.

Riker, W. H. and P. C. Ordeshook (1968): "A Theory of the Calculus of Voting," *The American Political Science Review*, 62, 25–42.

Roth, A. E. (2007): "Repugnance as a Constraint on Markets," *The Journal of Economic Perspectives*, 21, 37–58.

Shayo, M. and A. Harel (2012): "Non-consequentialist voting," *Journal of Economic Behavior & Organization*, 81, 299–313.

Sinnott-Armstrong, W. (2012): "Consequentialism," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta.

Smith, A. (1761): *The Theory of Moral Sentiments*, A. Millar.

Smith, K. D., J. P. Keating, and E. Stotland (1989): "Altruism Reconsidered: The Effect of Denying Feedback on a Victim's Status to Empathic Witnesses," *Journal of Personality and Social Psychology*, 57, 641–650.

Starmer, C. and R. Sugden (1991): "Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation," *The American Economic Review*, 81, 971–978.

Titchener, J. L. (1967): "Experimenter Effects in Behavioral Research," *Archives of Internal Medicine*, 120, 753–755.

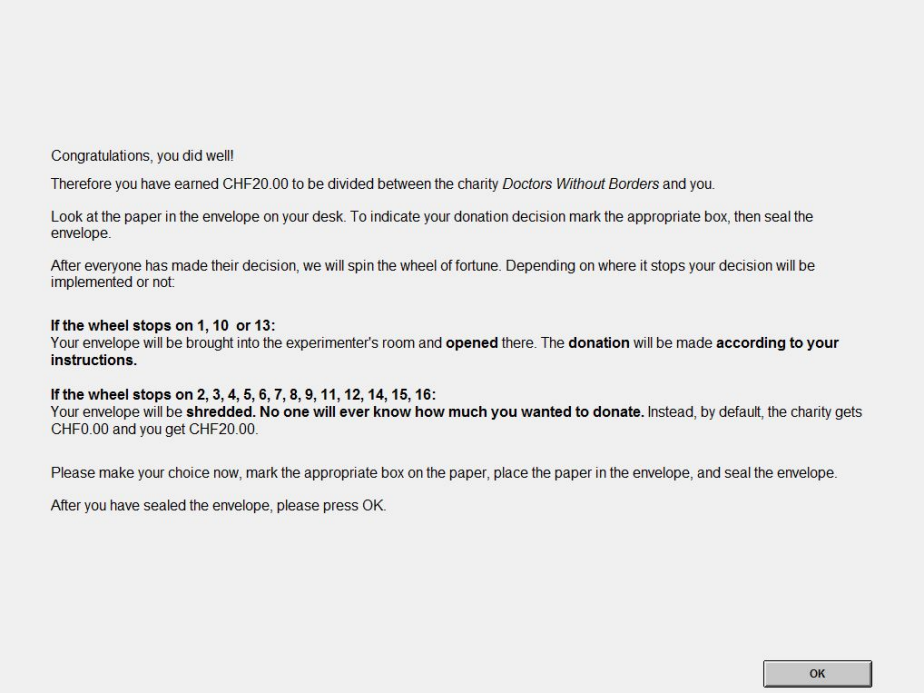
Tversky, A. and D. Kahneman (1992): "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, 5, 297–323.

Tyler, T. R. and Y. J. Huo (2002): *Trust in the Law: Encouraging Public Cooperation with the Police and Courts*, Russell Sage Foundation Series on Trust, Russell Sage Foundation.

Wilcox, N. T. (1993): "Lottery Choice: Incentives, Complexity and Decision Time," *The Economic Journal*, 103, 1397–1417.

Web Appendix:

Shredding Experiment Instructions Donation Screen for Subject with $\pi = 3/16$ and $\kappa = 0$



Sheet of paper participants fill out, put in an envelope, and seal.

Donation decision of subject number: 2

If you see the congratulations screen**:**

Of the CHF20 I want to donate

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

CHF to Doctors Without Borders.

If you have made too many mistakes**:**

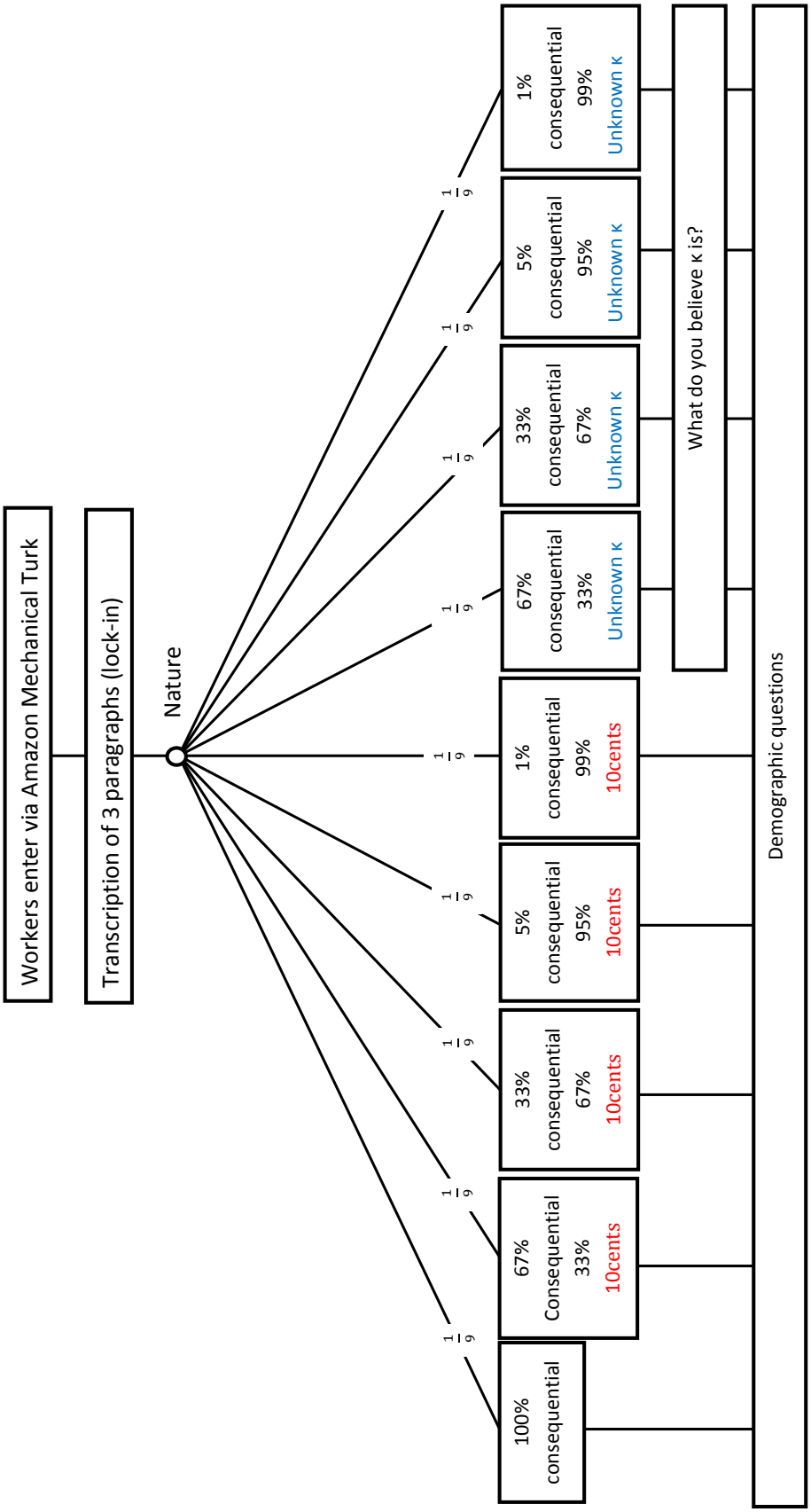
Please check this box: ☐

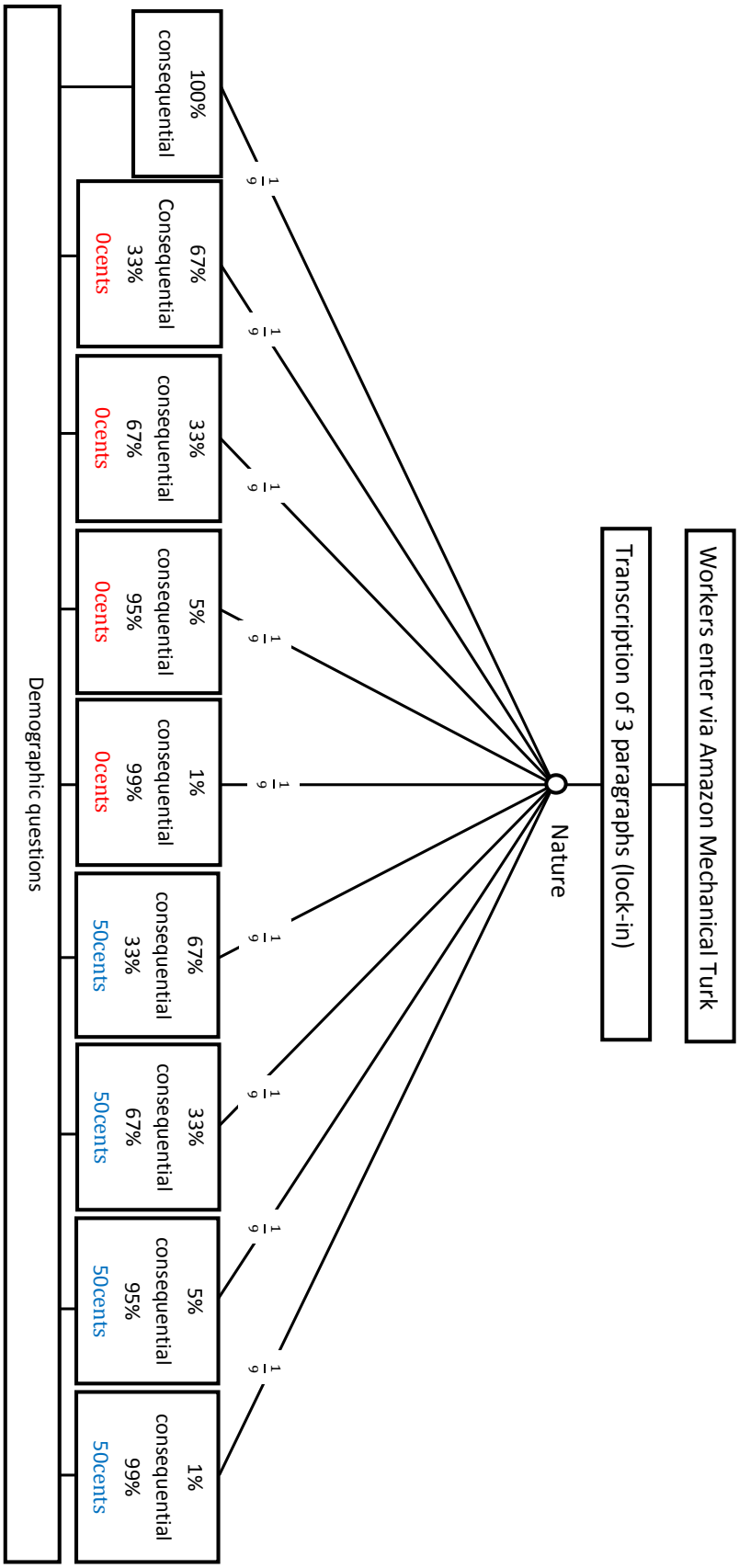
After marking exactly one box, please put this sheet in the envelope and seal it.

→ Then click OK on the screen so the experiment can proceed!

As *DM* becomes less consequential, what happens to the donation? (=What is the sign of $-\frac{\partial d}{\partial \pi}$?)						
Experiment	Experimental evidence	Consequentialism	Purely Deontological	Targeting Ex Ante	Consequentialist-deontological	
					additive, each concave	general
Red Cross κ=ω=50	+	0	0	-	+	-+
Red Cross κ=0	+	0	0	+	+	-+
Red Cross κ=10	+	0	0	?	+	-+
Red Cross κ=unknown	+	0	0	?	+	-+
Co-worker κ=unknown	-	0	0	?	+	-+

Note: Self-signalling can be interpreted as consequentialist, deceptive self-signalling as consequentialist-deontological.


Workers enter via Amazon Mechanical Turk
Transcription of 3 paragraphs (lock-in)
Nature
1/9
100% consequential
1/9
67% Consequential 33% 10cents
1/9
33% consequential 67% 10cents
1/9
5% consequential 95% 10cents
1/9
1% consequential 99% 10cents
1/9
67% consequential 33% Unknown κ
1/9
33% consequential 67% Unknown κ
1/9
5% consequential 95% Unknown κ
1/9
1% consequential 99% Unknown κ
What do you believe κ is?
Demographic questions


Workers enter via Amazon Mechanical Turk
Transcription of 3 paragraphs (lock-in)
Nature
1/9
1/9
1/9
1/9
1/9
1/9
1/9
1/9
1/9
100% consequential
67% Consequential 33% 0cents
33% consequential 67% 0cents
5% consequential 95% 0cents
1% consequential 99% 0cents
67% consequential 33% 50cents
33% consequential 67% 50cents
5% consequential 95% 50cents
1% consequential 99% 50cents
Demographic questions