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 an 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. We show that 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 gives insight into the location of the optimand for one's duty. We provide a formal interpretation of major moral philosophies and a revealed preference method to detect deontological motivations and discuss the relevance of the theory and method for economics and law.

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, Toulouse Institute for Advanced Study, daniel.li.chen@gmail.com; Martin Schonger, Law and Economics, ETH Zurich, mschonger@ethz.ch. First draft: May 2009. Current draft: July 2016. Latest version available at: http://users.nber.org/∼dlchen/papers/Social_Preferences_or_Sacred_Values.pdf. We thank research assistants and numerous colleagues including Jennifer Arlen, Andrew Caplin, Alain Cohn, Florian Ederer, Ernst Fehr, Oliver Hart, Larry Katz, Michael Kremer, Michel Marechal, Jean-Charles Rochet, Moses Shayo, Rani Spiegler, Uri Simonsohn, Bertil Tungodden, Alex Wagner, Roberto Weber, Georg Weizsäcker, and Matthias Wibral 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 (Grant No. 614708), Swiss National Science Foundation (Grant Nos. 100018-152678 and 106014-150820), Ewing Marion Kauffman Foundation, Institute for Humane Studies, John M. Olin Foundation, Agence Nationale de la Recherche, and the Templeton Foundation (Grant No. 22420).

1. INTRODUCTION

There is a classic divide between the consequentialist view that optimal policy should be calculated from considerations of costs and benefits and an alternative view held by many non-economists that policy should be determined deontologically (Harsanyi 1977; Binmore 1994; Smith 2016). That is, people, society, and judges, have certain duties and from these duties we derive what is the correct law – what is right and just. This paper does not take a stance on one view or the other. Rather, we ask the behavioral question: Are there deontological motivations, and if there are, how might we formally model these motivations? What are the implications of things like deontological motivations for economics, or what puzzles can we explain with deontological motivations that we cannot with standard models?

In the last few decades, there has been a gradual expansion of the domain of preferences considered by theory. There is the homo-oeconomics view – that people are only motivated by material consequences of their decisions on their own payoffs. Confronted with mounting lab evidence that people are consistently nice (pro-social) in economic games, the models were expanded under a broader rubric of incorporating fairness into economics (Rabin 1993), so that people care about the consequences of their decisions for others; that is, I care about inequality between my own payoffs and the payoffs of others (Fehr and Schmidt 1999). More recently, models have been expanded to incorporate what do people think of my type or my intentions (McCabe et al. 2003; Falk and Fischbacher 2006; Bénabou and Tirole 2006) and the social audience for one's decision (Andreoni and Bernheim 2009). Just knowing that the receiver will think badly of the decision-maker can be sufficient to compel giving (Dana et al. 2006, 2007). This paper considers another slice, that is, can people be motivated simply out of duty – purely internal consequences? That isn't to say that homo oeconomics, inequity aversion, and social audience motivations aren't relevant when we observe pro-social behavior. We will also be unable to isolate deontological motivations, that is, we will need deontological motivations and another behavioral assumption to explain observed behavior. What the paper presents is a thought experiment and a theorem that predicts invariance in the experiment–if people are motivated solely under the first three sets of motivations–but, if the fourth motivation is present (and in combination with at least one of the first three), then this thought experiment and actual experiment will predict variance.

In general, consequentialism and deontological motivations are very hard to distinguish. Often we may ask people, why did you make that decision, and they may say it's motivated out of duty or it's the moral or just thing to do. There's also work by experimental philosophers involving vignettes, like the moral trolley problem (Greene et al. 2001; Mikhail 2007), and experimental economics protocols ranging from cooperation (Rand et al. 2012), to truth-telling (Gneezy 2005), to saving the lives of mice (Falk and Szech 2013), as well as protocols involving priming with moral arguments (Cappelen et al. 2013a). What we propose is a method using revealed preference to identify non-consequentialist motivations. To preview, the thought experiment is to vary the probability that the decision is implemented.

We can think of the thought experiment as coming from Kant. In a classic vignette, your friend

is hiding in your house from a murderer. The murderer arrives and asks you whether your friend is hiding in your house (Kant 1797). Assuming you cannot stay silent, should you lie or tell the truth? So in the categorical imperative, Kant would say, "You must not lie." No matter what the consequences are, you must tell the truth. He allows for uncertainty–the possibility that your decision has the ultimate adverse consequence or has no consequences[1]–but "to be truthful in all declarations is a sacred and unconditionally commanding law of reason that admits no expediency whatsoever."[2]

We can put an abstract form to the categorical imperative. Think of a decision-maker (DM) making a decision d. We want to separate the motivation for the decision from the motivation for its consequences. Consequences can be broad, including reputation, inequality, and own payoffs. Consequences x is a function of the state of nature and decision d. There are two states, in the consequential state, d becomes common knowledge and is implemented. In the non-consequential state, d remains unknown to anyone, including the experimenter. With consequentialism, preferences are over lotteries (Anscombe and Aumann 1963). With deontological motivations, d matters *per se*, even in the non-consequential state.

Think of d^1, d^2, .., $d^{|D|}$, as possible decisions (Figure 1). With some probability, π, your decision is implemented–has consequences–and with $1-\pi$, your decision has no consequence. So x^C is a function of the decision and x^N, some constant outcome that's invariant to your decision. This thought experiment can apply to any decision with a moral element, but in a dictator game (Figure 2), you have your endowment ω, and you can donate anywhere from 0 to ω; with some probability π, decisions are carried out. The recipient receives d and you receive the $\omega - d$. With probability $1-\pi$, your decision is not implemented; recipient receives κ and you keep the remainder.

We can do this online to have large samples for structural estimation of the optimand of one's duty or in the lab with a public randomization device that makes the decision consequential with some probability (Figure 3). Subjects are asked to put their decisions in sealed envelopes, and with some probability, their envelope is shredded, so it has utterly no consequences, *not even through the experimenter*. So we eliminate even that element of social audience (Cilliers et al. 2015) or social altruism in the societal good effectuated by providing one's data for science. With the other probability, the envelopes are opened and the decision has consequences to be carried out.

The closest analog in the field may be the decision to sign-up as a bone marrow donor. Bergstrom et al. (2009) find that ethnic groups that are more likely to be called off the list to donate are less likely to sign up to be a bone marrow donor. With some probability your decision has consequences and the recipient receives the bone marrow while you undergo expensive and painful surgery. The decision to sign-up would be analogous to d, except the decision is not anonymous. Another analog may be the decision not to abort a fetus with Down Syndrome, and American subjects are asked, as prospective parents, as parents diagnosed as high-risk, and as parents whose fetus has already

[1]"It is indeed possible that after you have honestly answered Yes to the murderer's question as to whether the intended victim is in the house, the latter went out unobserved and thus eluded the murderer, so that the deed would not have come about."

[2]More recently, in a literature on why individuals obey the law, perceived legitimacy motivates obedience to rules irrespective of likelihood of reward or punishment (Tyler 1997).

FIGURE 1.— Thought Experiment

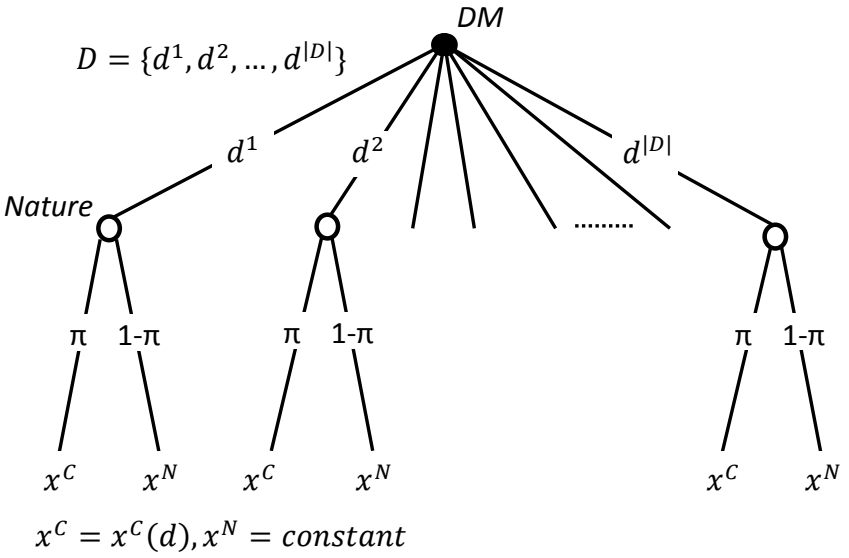


FIGURE 2.— Duty to Donate

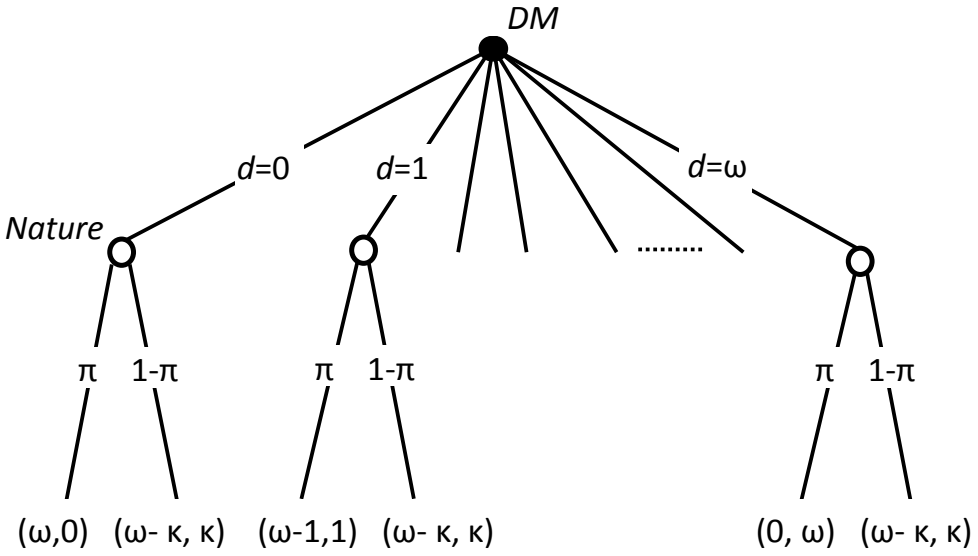
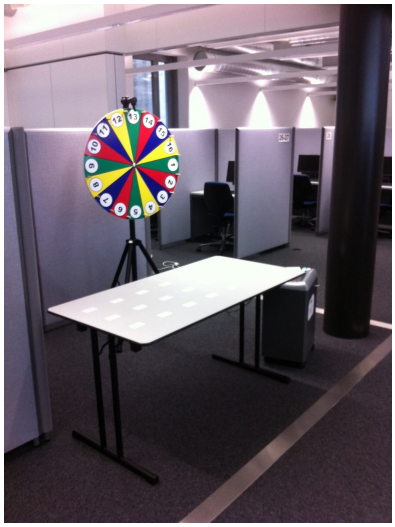


FIGURE 3.— Lab Implementation



been diagnosed (Choi et al. 2012). How does the decision vary with the probability of it being consequential? Other field analogs are challenging because changing the probability of implementation may affect the perceived worth of the decision or its beneficiary (in our experiments, the beneficiary is the Red Cross and Doctors Without Borders).

It's straightforward to show that pure deontologists following the categorical imperative would not change their behavior as the probability changes, but it turns out that pure consequentialists also do not change their behavior. One way to see this is with expected utility. For the decision-maker donating the marginal penny, the marginal benefit of donating is the recipient's well-being and any consequence of that increase. The marginal cost is to give up that penny. The decision-maker equates the marginal benefits and marginal costs. If the probability that this decision is implemented falls, then both the marginal benefits and costs fall, and they fall equally. So the decision-maker still makes the same decision on the margin. Formally, the indirect objective function is proportional to the utility of the decision implemented with certainty. So the decision is unaffected by π.

We can significantly relax the assumption of expected utility. One only needs first order stochastic dominance to obtain invariance. That is, someone who satisfies the behavioral assumption of first order stochastic dominance (FOSD) and is purely consequentialist–any or all of the first three sets of motivations–will not change their behavior as the probability changes. We provide a graphical and formal proof. The theorem is not Savage's (1972) Sure-Thing Principle. In Savage (1972), if the decision is optimal if the Republican candidate wins and is optimal if the Democratic candidate wins, then the decision is invariant to the probability of whichever candidate wins. In Machina and Schmeidler's (1992) formulation of the Sure-Thing Principle, the decision is invariant to κ. First order stochastic dominance is also not the independence axiom (for example, cumulative prospect theory and rank dependent expected utility satisfy first order stochastic dominance but not independence).[3] Formally, our theorem needs both strict FOSD and weak FOSD since the former does not imply

[3]FOSD is similar to Savage's (1972) behavioral postulate of Eventwise Monotonicity.

the latter; Machina's (1989) mom with an indivisible treat to allocate to one of her two children is a counterexample.

Of course, like Machina's mom, the DM may wish to target the ex ante utility of individuals, whether the expected income of the recipient or the expected giving by the DM. Such a DM, would donate more as the probability of her decision being consequential declines. We formalize the potential confound and design a treatment arm where the non-consequential state of the world involves the entire sum being donated to rule out ex ante consequentialism as motivation. Another possible explanation for variance 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 non-consequential state is realized. We also formally model and experimentally test cognition costs as a potential confound in our setting. For a previous formalization, albeit one that does not have the decision-maker solve the metaproblem optimally, see Wilcox (1993).

We illustrate the theory with three empirical exercises: an online experiment, a lab experiment, and structural estimation. Each analysis offers unique advantages and limitations that taken together, along with the observational evidence, portray a picture of variance in response to the probability of implementation. The first experiment uses Mechanical Turk, where large samples allows very low implementation probabilities. A linear regression model indicates that subjects became 33% more moral when the decision was hypothetical. The second experiment uses the lab. Subjects became 50% more moral when the decision was hypothetical. Structural estimation calculations suggest that behavior that appear as consequentialist Fehr-Schmidt preferences may be largely explained by deontological motivations in our experiment under a bliss point functional form assumption.

The results were not due to subjects targeting an expected income of the recipient nor to targeting the expected giving by the subject. The differences were statistically significant with linear regressions and non-parametric tests. The magnitudes are comparable to field analogs: the decision to sign-up to be a bone marrow donor is 3.5 times more likely, as the probability your decision is implemented falls (from .3 to .1%) (Bergstrom et al. 2009) and prospective parents are 14 times more likely to keep the fetus as the probability the decision is implemented falls from 100% to .1% (Choi et al. 2012). We use a between-subject design to avoid sequence effects, but heterogeneity among subjects would introduce noise, biasing our estimates to 0, so long as there is random assignment and no attrition conditional upon treatment being revealed in the online setting.

The remainder of the paper is organized as follows. Section 2 reviews the literature. Section 3 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 in our thought experiment. Section 4 describes the actual experiments along with the results. Section 5 presents the structural estimation. Section 6 discusses the results and Section 7 concludes.

2. LITERATURE REVIEW

> "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." (emphasis added) (Smith 1761).

Adam Smith's (1761) impartial spectator in The Theory of Moral Sentiments may have been duty-oriented though perhaps also consequentialist. In more recent literature, non-consequentialist motivations can be viewed as a subset of motivations such as warm glow (Andreoni 1990), self-image (Bénabou and Tirole 2011), lying aversion (Gneezy 2005), expressive voting (Riker and Ordeshook 1968), and distributive justice (Elizabeth Hoffman 1985). In a seminal theoretical contribution, Andreoni (1990) points out that DMs in a public goods contribution framework can derive utility not only from the total amount of the public good G provided, but also from her contribution g. However, Andreoni and Bernheim (2009) points out that social audience is a possible micro-foundation for warm glow: the effects of g may be channeled through what others observe, what the experimenter observes, or through the receipt itself, rather than the decision. In a recent theoretical contribution, 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. Similar to the flexibility in Andreoni's framework, DMs may care about the fact that the decision is implemented rather than the decision itself and the verbal description of "identity investment" resonates with a consequentialist interpretation. Gneezy (2005) and subsequent work provide a fascinating lying aversion experiment that varies the cost of saying the truth. However, saying the truth has consequences to observers and to payoffs (which, incidentally, is why we use a donation game rather than a truth-telling game). In another important insight, people may participate in elections even when their vote is not pivotal due to a "duty" to vote (Riker and Ordeshook 1968). Feddersen et al. (2009) and Shayo and Harel (2012) formalize the insight that individuals may have expressive motivations, that is, they obtain a small positive payoff by the act of voting for an option independent of the electoral outcome, which they test with experiments varying the probability of being pivotal. However, expressive motivations can include voting to tell others (DellaVigna et al. 2013), sending a message to the public, and affecting the extent of a candidate's mandate, since the election outcomes are public. These other motivations resonate with a consequentialist interpretation. Finally, distributive justice (Elizabeth Hoffman 1985; Konow 2000) is another motivation related to deontological motivations but includes consequentialist ones.

It is difficult to do justice to the large number of important contributions on non-payoff outcomes (as opposed to non-consequentialist) based motivations. To name a few: Ellingsen and Johannesson (2008) has a utility function incorporating own payoff, others' payoff, and how others think of me, but does not have the action directly, except via these consequences. Battigalli and Dufwenberg

(2007) model guilt aversion; the prototypical cause of guilt would be the infliction of harm or distress on the recipient. More broadly, psychological games involve preferences over the beliefs of others. In our setting, we rule out a psychological game of consequences in response to the probability of implementation by keeping the decisions anonymous.[4] Preferences for procedural fairness (Gibson et al. 2013; Brock et al. 2013) are important motivations related to but distinct from deontological ones.

In a research design that is close but includes subtle differences, Andreoni and Bernheim (2009) use a modified dictator game with random implementation probabilities. It differs in five dimensions. First, we make both the probability and the realization of the state of nature public; in their study, dictators can hide their selfishness behind nature's choice and the fact that nature chose the outcome is not observed by recipients. Second, we make the recipient a charitable organization outside the lab; in their study, the recipients are in the room and dictators become *more* generous as the probability of implementation increases because they are motivated by their social audience. Third, the lab experiment shreds decisions; in their study, they view motivations regarding what the experimenter infers when the decision is not implemented as a confound. Fourth, we do not use the strategy method; in their study, subjects play several games, reducing the likelihood a particular decision is consequential, whereas in our study, each subject sees only one probability. Fifth, making one decision minimizes sequence effects, and they recognize the importance of not using within-subject variation for any particular game. For example, if an experimenter asks two questions with a higher and lower probability, subjects may feel the right answer is to give more in one question, which would be a confound for our invariance theorem.[5]

We can also contrast our use of random implementation probabilities with a large literature in psychology that varies the probability that one's help will have an impact conditional on experiencing the cost of helping (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*: the cost of the decision is experienced by subjects whether or not their decision to help is effective. These studies find, like Andreoni and Bernheim (2009), that as the probability falls, generosity declines, while we find the opposite.

In sum, none of the related papers formalize major moral philosophies nor focus on deontological motivations *per se*, which seem to be important in many contexts, in law, in religion, and in a conflict of "sacred values" (Bowles and Polania-Reyes 2012) and repugnance (Roth 2007; Mankiw and Weinzierl 2010).[6] If our theorem is correct, to deny the variance of decisions as the probability of im-

[4]Therefore, the confound would only involve a psychological game with respect to the experimenter, but even then, subjects have limited interactions with the experimenter in the lab–at the time of payment–or online, where there is effectively zero interaction.

[5]Grossman (2015) also uses a modified dictator game with random implementation probabilities, but each participant played the role of dictator and served as recipient for someone else, which can lead to strategic considerations regarding beliefs about other players.

[6]Motivations like honor (Nisbett and Cohen 1996) or duty may be related to conflicts of "sacred values": this may apply to fundamentalisms (Chen and Lind 2007, 2014; Chen 2006, 2010) or political polarization (Berdejó and Chen 2014). Legal compliance may be driven in part by perceived legitimacy of law: for instance, in the duty that minorities may feel to comply with the law (Tyler and Huo 2002; Chen 2013). The U.S. Department of Defense has begun an

plementation falls would deny the existence of deontological motivations (or FOSD). Deontological motivations are relevant for economics. If responses differ by 33 to 50% when decisions are hypothetical, then our results support the use of consequential decisions made under revealed preference rather then contingent valuation and psychologists' use of vignettes. Ours is a formal decision-theoretic observation that supplements non-formal psychological observations (List and Gallet 2001; List 2001; Diamond and Hausman 1994). However, unlike the non-formal statements regarding vignette studies, an implication of our model is that any probability between 100%-consequential and 0%-hypothetical can affect decision-making when moral considerations are relevant. Thus, one framing of existing methods like the random lottery method where many decisions are elicited but only one is carried out is: they may reveal decisions that are more moral, and treatment effects may be larger and more statistically significant (Bertrand et al. 2004) than if the decisions were consequential. Another framing is that these methods including the strategy method incidentally turn out to help understand the general way in which agents' motivations influence behavior (Camerer 2011) insofar as the reduction in the likelihood of the decision being consequential reveals the optimand of one's duty.

Some formal observations support the use of the random lottery method and the observations have been empirically validated (Starmer and Sugden 1991; Hey and Lee 2005). Roughly speaking, if individuals satisfy the independence axiom (Holt 1986), then the random lottery method is valid. In contrast, we show if individuals satisfy a behavioral postulate–FOSD–weaker than the independence axiom or have linear utility, the random lottery method reveals different decisions when moral (deontological) decisions are at stake. We distinguish the prior meta-analyses of economic experiments supporting the irrelevance of the probability by making a distinction between games with purely consequentialist motivations (such as in games simulating firms) vs. those potentially with non-consequentialist motivations (such as in social preference games); differences emerge in a literature review when focusing on the latter (Chen and Schonger 2015).

An example where deontological motivations may be relevant is the concept of *intent* in law, most famously, in criminal law when a distinction is made between *mens rea* (intention) and *actus reus* (act): did the shooter intend to kill but did not kill *or* did the shooter unintentionally kill, is a common hypothetical for law students to ponder. In many other instances, the law cares about mental states beyond just the consequences: Judges are sometimes interested in the motivations of the litigant in copyright disputes, where a litigant has cause of action only if she is motivated by her *moral rights* to litigate, that is, she is not litigating because of the consequences of winning. In equity law, judges may care about opportunistic behavior, which is similar to having both *mens rea* and *actus reus.* For philosophers who argue that human dignity derives from the possibility of deontological decision-making (in the sense that the decision is not coerced in any way

initiative to understand and change sacred values. The difficulty of "winning the hearts and minds" through law and markets and of shifting injunctive norms with formal institutions is an open question (Chen and Yeh 2014a; Chen 2014a,b; Chen and Yeh 2013). Welfare economics or policy responses that incorporate deontological motivations are relevant in how government policies affect doctors' duty to care (Chen et al. 2014) or whether policy change should be gradual or sharp when anticipating backlash (Chen 2004; Chen et al. 2014).

due to its consequences) and for those who stress the importance of screening for the presence of deontological motivations in business leaders, politicians, or judges (Besley 2005), methods such as ours may be relevant. Returning to the motivating debate, recent theoretical work criticizes the non-consequentialist (non-welfarist) approach to optimal policy design as necessarily harming some individuals (Kaplow and Shavell 2006, 2001). If people are proven to have non-consequentialist motivations, then the divide between the different approaches to optimal policy design may be less than previously thought.

Beyond the scope of this article are questions of how deontological commitments emerge, their persistence throughout a human life, and how they are shaped by culture, education, and policy. Someone may hold a deontological view because of the consequences of that view, e.g., rule consequentialism. A related example for this is a moral decision-maker, "homo kantiensis", whose preferences are ones that are socially optimal when everyone else also holds that view (Alger and Weibull 2013). Alger and Weibull (2013) report that such preferences are selected for when preferences rather than strategies are the unit of selection and they find that preferences that are a convex combination of homo oeconomicus and homo kantiensis will be evolutionarily stable. Our framework has little to say about the source of–or reasons for–deontological commitments; rather, it asks, at the moment of making a decision, is the DM deciding deontologically holding apart the consequences.

3. THEORY

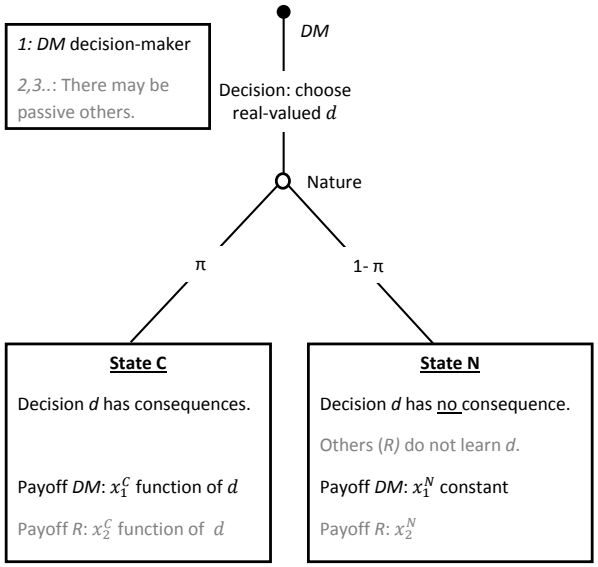
We introduce our thought experiment and focus on consequentialism and the invariance theorem first. We illustrate the intuition for the theorem under expected utility (this intuition is a corollary of the main theorem), a graphical proof of the invariance theorem, and then the formal statement of the assumptions along with the theorem itself. Next, we formalize deontological motivations as a lexicographic preference–duty first, then consequences–and show invariance still holds. We then show variance when individuals have both consequentialism and deontological motivations and the direction of change under additive separability.

3.1. *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 (x_1^N, x_2^N) in state N. Figure 4 illustrates this.

This general experimental design could be used for many morally relevant decisions; here we

FIGURE 4.— Thought Experiment: General Idea



apply our identification method to the dictator game and thus to the moral decision to share. As shown in Figure 5, 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.

3.2. *Intuition*

We illustrate the intuition of the invariance theorem under expected utility. Given expected utility, the DM maximizes:

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

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

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

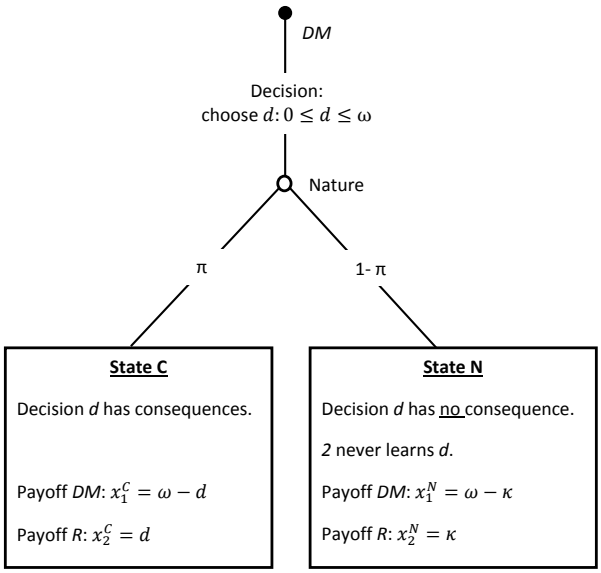
Limiting attention to pure consequentialists, the problem simplifies to:

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

and the indirect objective function to:

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

FIGURE 5.— Actual Experiment



Note that now the d does not enter in the second term, which corresponds to state N. The indirect objective function is proportional to $u(\omega - d, d)$, so $\frac{\partial d^*}{\partial \pi} = 0$.

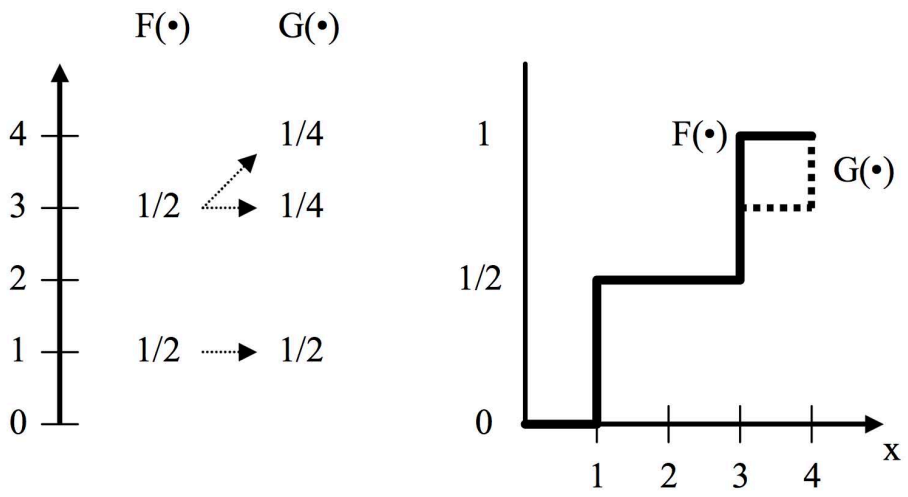
3.3. *Graphical proof*

In the previous subsection, we have seen that if the DM satisfies the axioms of expected utility then if d^* is not constant in the probability she cannot be a consequentialist. Put differently, if we observe a DM to vary her decision in the probability we would reject the joint hypothesis that the DM is a consequentialist and an expected-utility maximizer. Since expected utility theory often fails to describe behavior (Starmer 2000) such a joint test would tell us little about whether consequentialism or expected utility or both were rejected. It is therefore desirable to have much weaker assumptions about decision-making under objective uncertainty than expected utility theory. Here we show that first-order stochastic dominance is sufficient for the result.

First, we provide a graphical sketch of the invariance proof. That is, someone who satisfies the behavioral assumption of preference relations over FOSD and is purely consequentialist will not change their behavior as the probability changes. The left-hand side of Figure 6 provides an example of FOSD. Think of an ordering over outcomes, 0, 1, 2, 3, and 4 on the Y-axis and the corresponding lotteries F and G. G looks better than F since instead of getting 3, sometimes you get 4. Formally, G first-order stochastically dominates F with respect to $\succsim$ if for all x': $\sum_{x:x' \succsim x} G(x) \leq \sum_{x:x' \succsim x} F(x)$.

For every outcome x', the probability of any outcome worse than x' is lower under G than under F. That can be represented graphically on the right, as a CDF. For the proof, recall that decisions are choices over lotteries like F and G. Suppose 1 is the non-consequentialist outcome, and let 3 or 4 be the active choice. So in the experiment, there is no kink in the dotted line in the right-hand figure. The dotted line would extend at the y-axis at 1/2 (rather than 3/4) from 3 to 4. What does changing the probability do? It moves the horizontal bar up and down. But G always FOSD F. So

FIGURE 6.— First Order Stochastic Dominance



if a choice is optimal for one probability, it is the optimal choice for all probabilities.

3.4. *Formal statement of assumptions and theorem*

In the following, we show 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.

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. It is important to define FOSD with respect to ordering over outcomes rather than the outcomes themselves. FOSD over outcomes 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. Therefore, we define FOSD over the ordering of outcomes.

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 $p \succsim q$.*

Formally, our theorem needs both strict FOSD and weak FOSD since the former does not imply the latter.

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) $x \gtrsim x'$: 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 $x \gtrsim x'$, $\pi' x + (1-\pi')x''$, 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 others' 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.

3.5. *Examples*

The following examples are intended to distinguish deontological motivations from previous models of economic behavior, including social preferences, but some theoretically-inclined readers may wish to skip these examples. However, we will use the second example in our structural estimation. 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 inequity aversion:

EXAMPLE 2 **(Inequity aversion)** 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]

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

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 warm glow motivations:

EXAMPLE 3 (**Warm glow**) 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 Andreoni and Bernheim (2009) propose social audience motivations as a possible micro-foundation for warm glow. Then, a DM with warm-glow preferences may be a consequentialist whose preferences depend on her own outcome, the charitable recipient's outcomes, and the decision that is implemented. 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]

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

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

The indirect objective function is then:

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

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 or the decisions carried out. 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).

3.6. *Strict FOSD does not imply FOSD*

Formally, our theorem needs both strict FOSD and weak FOSD since the former does not imply the latter. We provide a counter-example and then the conditions under which strict FOSD implies FOSD, but some empirically-oriented readers may wish to skip these conditions. However, we design our experiment around a confound suggested by the counter-example.

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 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 $(x; \frac{1}{2}, y; \frac{1}{2}) \succ (x; \frac{2}{3}, y; \frac{1}{3})$.

Are there assumptions besides strict FOSD so we don't need both strict and weak FOSD?

3.6.1. *Continuity*

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 $\{\alpha \in [0,1] : x \succsim \alpha x + (1-\alpha)y\} = [0; \frac{1}{2}) \cup (\frac{1}{2}, 1]$ 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.

3.6.2. *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: $\sum_{x \precsim z} k(z) = 1 = \sum_{x \precsim z} l(z)$.
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.

3.7. *Potential Confounders*

A potential confound to testing the invariance theorem in an experiment 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 the famous example articulated by Machina (1989) and recapitulated above 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 5 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:

$$\frac{1}{\pi}\frac{\partial^2 V(d)}{(\partial d)^2} = -\left(2f_{12}f_1f_2 - f_{11}f_2^2 - f_{22}f_1^2\right)\frac{1}{f_2^2}\pi u_1^2(\omega - d) + f_1 u_{11}(\omega - d) + f_2 \widetilde{u}_{11}(d) < 0$$

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:

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

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. In sum, targeting the recipient's expected income can be assessed by our research design by seeing if the sign of $\frac{\partial d^*}{\partial \pi}$ flips in the two treatment arms. Motivations pertaining to forms of residual uncertainty that take into account ex-ante considerations but mix them with ex-post considerations would also predict the sign to flip.

3.8. *Defining consequentialism and deontic motivations*

Before presenting our formal interpretation of the philosophical concepts of consequentialist and deontological moral philosophy, let us review their definitions. Readers familiar with the philosophy may skip this section, which is based on *The Stanford Encyclopedia of Philosophy*. 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). Virtues ethics, which originates in the work of Plato and Aristotle, would also be among the non-consequentialist motivations we seek to uncover. To economize on terminology, we will only refer to deontological ethics. We also make no distinction between positive and negative duties.

In our 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; Binmore 1994). We do not take a normative stance on these issues, rather we take the behavioral question of whether people have deontological motivations and how might we model these motivations. 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 that objective function and refrain from interpreting the function as standing for utility or happiness.

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, 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 subsections is that we extend the domain of preferences 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. The standard theories of decision-making by Savage (1972) and Anscombe and Aumann (1963) rely on the

assumption that the domain of consequences is state independent.[9]

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 mirroring the 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.

It is possible to model purely deontological people as having a different choice set (Nozick 1974). But traditionally a choice set is the objective, external constraints facing a person and 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. 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 (i.e., the optimand of one's *greatest* duty rather than the optimand of one's duty), 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).

[9]Evaluation of state-contingent outcomes is formal analyzed elsewhere (Chen and Schonger 2015).

DEFINITION 3 DEONTOLOGICAL PREFERENCES: A preference is called *deontological* if there exist u, f such that $u = u(d)$, and $f = f(x)$, and for all $(x, d), (x', d')$: $(x, d) \succsim (x', d')$ 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 $V(d)$.

It is a natural question to ask if deontological moral philosophy even applies to situations under objective uncertainty. Kant would say yes. The natural world is always uncertain, so any moral philosophy that aims to provide guidance to people in the natural world presumably must apply to decision-making under uncertainty. Let us illustrate this point with what Kant thinks about uncertainty in the famous ax-murderer example: He states that there is always some uncertainty about the consequences of saying the truth (or lying) to the ax-murderer; despite this, one should do one's duty to say the truth, regardless of the resolution of uncertainty and regardless of the consequences.

> "It is indeed possible that after you have honestly answered Yes to the murderer's question as to whether the intended victim is in the house, the latter went out unobserved and thus eluded the murderer, so that the deed would not have come about. [...] To be truthful (honest) in all declarations is, therefore, a sacred and unconditionally commanding law of reason that admits of no expediency whatsoever." (Kant 1797)

FACT 1 ***(Deontological preferences)*** *For purely deontological preferences the optimal decision d^* is constant in the probability π.*

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 $u(d)$.

3.9. *Consequentialist-deontological preferences*

Thus far, we have shown 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.

General Formulation

Consider the following minimal characterization of consequentialist preference.

$U(\kappa, d, \pi) = F(\kappa, \pi, u(d))$ such that for all κ and π, F is monotone increasing in u and weakly separable. Then no matter what the values of κ and π, a subject who chooses d to maximize U will choose the same d, namely the one that maximizes $u(d)$.

A mixed consequentialist-deontologist utility can be written as

$U(\kappa, d, \pi) = G(d, \kappa, \pi, u(d))$ such that the best choice of d depends on κ and/or π. For example, it might be that you get a good feeling about pledging d regardless of whether you have to do it or not. So a simple example would be

$G(d, \kappa, \pi, u(d)) = ad + (1-\pi)u(\kappa) + \pi u(d)$.

In this case an interior maximizing choice satisfies $u'(d) = -\frac{a}{\pi}$ and if $u'' < 0$, as π increases, the chosen d decreases.

Next, we provide a simple example of an additive utility function that depends on the decision d and on only one consequence, the payoff for the DM herself:

EXAMPLE 6 $u = u(x_1, d) = x_1 + b(d)$, 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 $\frac{\partial d^*}{\partial \pi} = \frac{1}{b_{11}(d)} < 0$. The second order condition is $b_{11}(d) < 0$.

For a slightly more general example: 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, the second order condition yields: $\frac{\partial^2 V(d)}{\partial d^2} = \pi f_{11}(\omega - d) + b_{11}(d) < 0$. Applying the implicit function theorem to the first order condition yields: $\frac{\partial d^*}{\partial \pi} = \frac{f_1(\omega - d^*)}{\pi f_{11}(\omega - d^*) + b_{11}(d^*)} < 0$, since utility is increasing in its own outcomes and the denominator which is the second derivative of the indirect objective function is 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 $\frac{\partial d^*}{\partial \pi} = 0$.

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. *In our setup, the benefit of altruism is always there, but the costs are only incurred with probability* π. 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] Next, we consider a common utility form, bell-shaped utility functions, used in estimates of policy choices by politicians.

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 bliss point: $d^* = \lambda\omega \equiv d_c^*$. For a DM who is pure deontological ($\mu = 1$) there exists a bliss point $d^* = \delta \equiv d_d^*$. 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.

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. These ideas are explored further in modeling and testing the shape of the cost of taking actions that one disagrees with morally or politically (Chen et al. 2015a).

3.10. *Cognition Costs*

Another 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

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

framework since they are incurred during the decision and are a consequence that even arises if the non-consequential state is realized. 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).

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 continuous. 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 contrast, 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 heuristic 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.

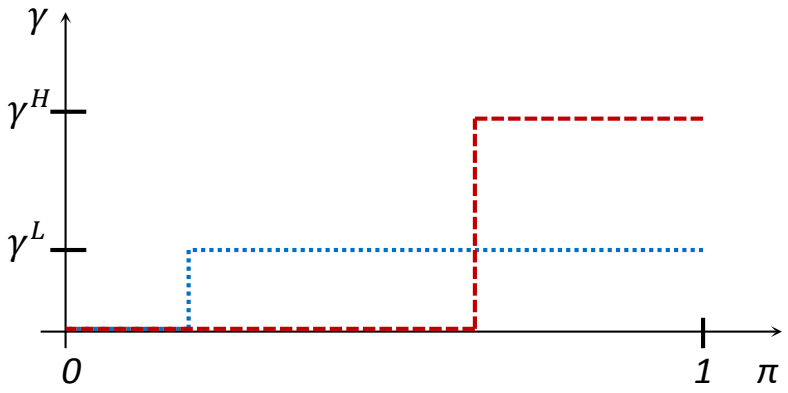
Now suppose we vary the cognition cost, that is, we do an exercise in comparative statics and investigate how $\widetilde{\pi}$ varies in γ, and note that

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

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. Figure 7 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

FIGURE 7.— S-Shape Cognition Costs



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 1) *time spent on the survey also changes with* π as d changes. Our research design provides a second test: 2) *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. Figure 7 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.

4. EMPIRICAL FRAMEWORK

We describe our empirical framework for testing variance with the probability of implementation across the three empirical settings.

4.1. *Specification*

Our baseline specification simply tests whether the decision is correlated with the treatment, conditional on a set of control variables:

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

π_i represents the probability of implementation for individual i. We execute a between-subject rather than within-subject design to minimize experimenter demand (Cilliers et al. 2015) and sequence effects (Chen et al. 2015b). In principle, controlling for individual demographic characteristics X_i is unnecessary with randomization. We display the raw data means, distributions, and results from non-parametric tests for differences in donations, and ordinary least squares regressions. β_1 measures the change in the donation as the probability of implementation increases. $\beta_1 < 0$ is ev-

idence in favor of increasing generosity in terms of mixed consequentialist-deontological motives. $\beta_1 = 0$ is evidence in favor of pure consequentialist or pure deontological motives. $\beta_1 > 0$ is evidence against us to the extent that the designated recipient was a salient charity, the Red Cross and Doctors Without Borders.

Recent observational studies (Bergstrom et al. 2009; Choi et al. 2012) mimic our experimental design and are suggestive of $\beta_1 < 0$. 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. The decision to sign-up to be a bone marrow donor is 3.5 times more likely, as the probability the decision is consequential falls from .3 to .1%. 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.[11] 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. Thus the likelihood of choosing to save the fetus increased by roughly a factor of 14 from 5% to 70% as the probability the decision is consequential falls from 100% to .1%.[12] These findings are similar to our thought experiment in that those whose actions are less likely to be carried out are also more generous. There is the important distinction that the decision is known to others, neither anonymous nor placed in an envelope.

4.2. *Experiment 1*

Experiment 1 was run on Amazon Mechanical Turk, so sample size was large. Workers came to the marketplace naturally and were unaware they were in an experiment at the time of arrival and were exposed to only one probability of implementation to minimize Hawthorne effects (Orne 1962; Titchener 1967). Through an interface provided by the MTurk, 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.

Online settings involve approximately costless exit from the experiment. Attrition is not problematic if the types of people who attrite are roughly identical across treatment conditions. To lock workers in and minimize attrition of different types of people in response to treatment, we first asked them to do data transcription of three paragraphs of text (Chen 2011; 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

[11] To be sure, since the setting is not an experiment, many other differences exist across ethnic groups, such as health, wealth, attitudes towards the medical profession, and probability of knowing someone who has needed a bone marrow transplant.

[12] This figure is based on roughly 1 in 700 births having Down Syndrome in the U.S.

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 $86.40 per day. The current federal minimum wage in the Unites States is $58/day. In India, payment rate depends on the type of work done, although the "floor" for data entry positions appears to be about $6.38/day.[13] In one data entry study, one worker emailed saying that $0.10 was too high and that the typical payment for this sort of data entry was $0.03 cents per paragraph.[14] Our decision-task involves reading essentially a single paragraph and making 1 decision, with an additional $0.50 possible, up to 17 times the expected wage.

Across many studies, the lock-in task is quite effective at minimizing attrition *when* treatment is revealed even when the subsequent experiment is very long (Chen and Yeh 2014a) and even when the attrition *before* treatment is revealed is as high as 25% in some cases (Chen 2011). After the lock-in task, subjects have an opportunity to split a bonus (separate from the payment they received for data entry) with the charitable recipient, the Red Cross.[15] 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 had 902 decisions from 902 subjects.

We ran the MTurk experiment twice. Both times, participants were randomly assigned to one of five groups with π being: 100%, 66%, 33%, 5%, and 1%, and told in advance 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 following the protocol described in the theory section (See Appendix Figure A.3). In a second experiment, we assess potential anchoring effects and 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 between 0 and 50. When κ was unknown, we also asked workers what they believed would be the amount donated if the computer made the decision (See Appendix Figure A.4). We present both the raw data as well as regression specifications that include indicator variables for κ. We also present the regression results disaggregated for each κ.[16] The results do not hinge on using the κ-unknown treatment arm.

Figure 8 shows the basic results: the lower the probability that the decision is consequential, the more generous is the decision-maker. The increase in generosity is monotonic with the decrease

[13]Payscale, Salary Snapshot for Data Entry Operator Jobs, http://www.payscale.com/research/IN/Job=Data_Entry-_Operator/Salary?, accessed June 17, 2011.

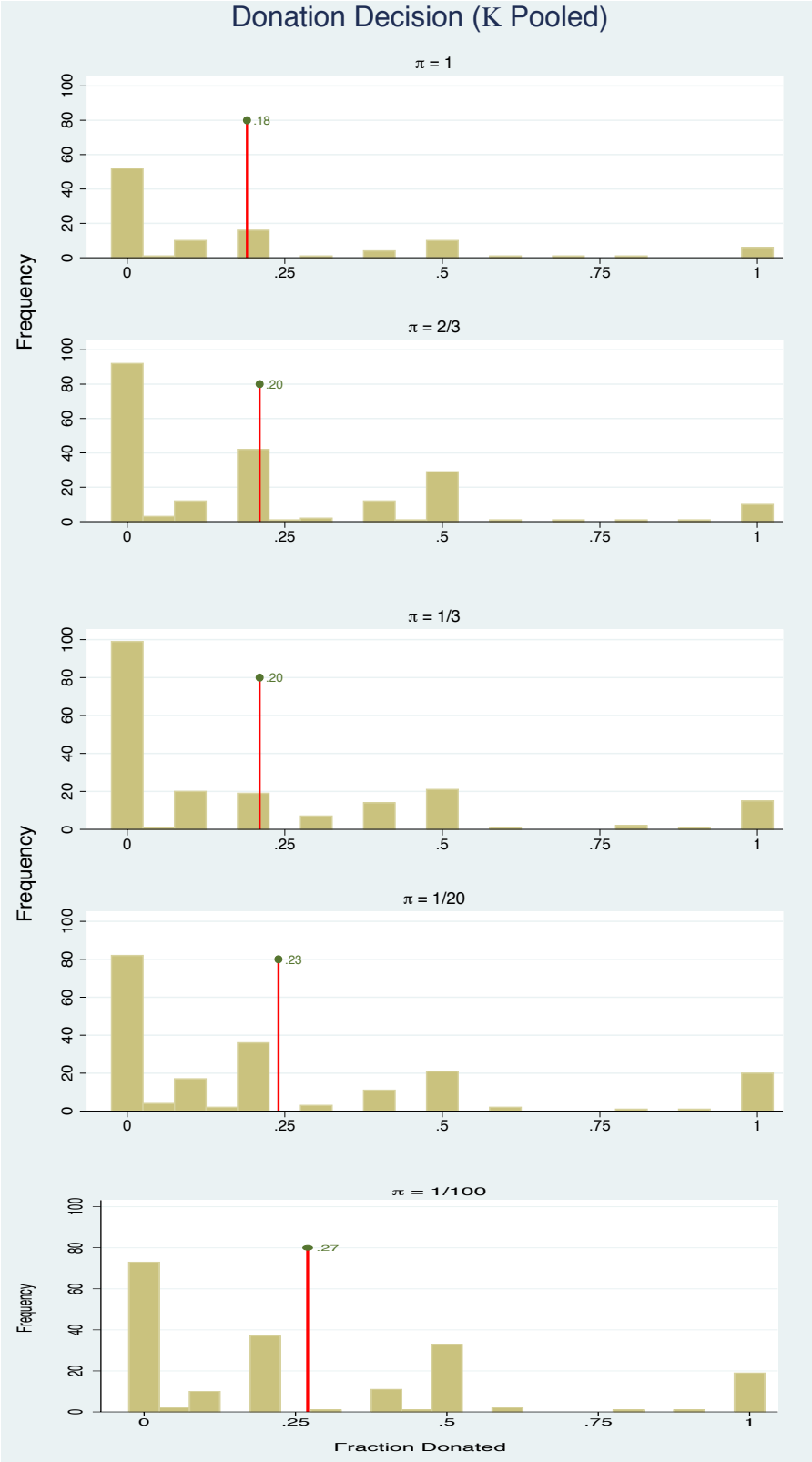
[14]A sample paragraph of data entry was a Tagalog translation of Adam Smith's *The Wealth of Nations*: Kaya sa isip o diwa na tayo ay sa mga ito, excites ilang mga antas ng parehong damdamin, sa proporsyon ng kasiglahan o dulness ng kuru-kuro.Ang labis na kung saan sila magbuntis sa kahirapan ng mga wretches nakakaapekto sa partikular na bahagi sa kanilang mga sarili ng higit pa sa anumang iba pang; dahil sa takot na arises mula sa kathang isip nila kung ano ang kani-kanilang mga sarili ay magtiis, kung sila ay talagang ang wretches kanino sila ay naghahanap sa, at kung sa partikular na bahagi sa kanilang mga sarili ay talagang apektado sa parehong miserable paraan. Ang tunay na puwersa ng mga kuru-kuro na ito ay sapat na, sa kanilang mga masasaktin frame, upang gumawa ng na galis o hindi mapalagay damdam complained ng.

[15]Most subjects are from the U.S. and India, and we believed the Red Cross to be more well-known in these countries. When asked to split the 50 cent bonus, they could report any integer from 0 to 50.

[16]In any event, potential anchoring effects are not significant: 18% of subjects gave 10 cents in the "$\kappa = 10$ Cents" treatment while 14% gave 10 cents in the "$\kappa =$ Unknown" treatment.

in probability. Donations increased from 18% (when $\pi = 1$) to 27% (when $\pi = 0.01$). Note that even if significant heterogeneity in the subject pool leads to variance with respect to the probability treatment, it's unlikely the idiosyncratic variation would vary monotonically with the probability.

FIGURE 8.— Donation and π: Raw data (MTurk)



Red: Mean

Table I reports that the effect of π is significant at the 5% level in a linear regression in Column 1. The effect size of 7.2% is roughly one-third of the mean donation of 23%. The R-square is a reasonably small 0.007 as can be seen from the dispersion in donations in Figure 8. Column 2 adds controls. 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. The point estimates are stable, while significance falls to the 10% level. The remainder of the table examines different versions of ex ante consequentialism. Columns 3 and 4 examine expected income and Columns 5 and 6 examine expected giving. We can strongly reject the hypothesis that subjects are targeting these quantities. 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%.[17]

TABLE I

Donation and π: Linear Regression (MTurk)

	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 10. 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 II presents separate linear regressions for each κ treatment-arm. In each pair of columns (without controls and with controls), 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.[18]

[17] To make calculations on expected donations when κ is unknown, we use data on perceived donation.

[18] We can also see that 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 II

DONATION AND π: LINEAR REGRESSION DISAGGREGATED BY κ (MTURK)

	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 distributions of donation decisions are significantly affected by π. Table III shows that along most thresholds for π, Mann-Whitney tests yield significant differences in the distribution of donations as π increases. To interpret, 0.05 in Column 1 means that we reject with 95% confidence the hypothesis that the distribution of decisions for subjects treated with $\pi = 1, 0.67, 0.33$ is the same as the distribution of decisions for subjects treated with $\pi = 0.05, 0.01$. The lower panel of Table III reports that the distribution of donations does not significantly vary by κ. Means are also not significantly different by κ.

TABLE III

DONATION AND π: NON-PARAMETRIC TESTS (MTURK)

	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	

4.3. *Experiment 2*

In second experiment, we ask lab subjects to make a donation decision with the knowledge that we will shred their decision when it is not implemented as the lab setting is less anonymous than the online setting. 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 (See Figure 3). The donation recipient was Doctors Without Borders as we believed this organization to be more salient than the Red Cross in German-speaking countries.[19] We ran the experiment in Zurich using zTree (Fischbacher 2007). In one session, we collected data from a classroom, where the endowment was 10Chf instead of 20Chf, so all our results are reported in terms of percentages from 0 to 1 (fraction donation) as in the previous sub-section.

We had a $2x2$ design: Subjects were 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 was spun, envelopes that were to be destroyed were collected and shredded. The remainder were opened and participants were paid. Among 264 subjects, 71 envelopes were opened. To minimize any strangeness subjects may have felt to be paid a large participation fee[20] and then only make one decision (which would probably be shredded), subjects were asked three IQ tasks analogous to the lock-in task to ensure they take the study seriously. If at least one answer was correct, they proceeded to the donation decision and received information about their probability of implementation (Appendix Figure A.1). The randomization wheel has sixteen numbers so we only need to mention one or three of these numbers to the subject. The

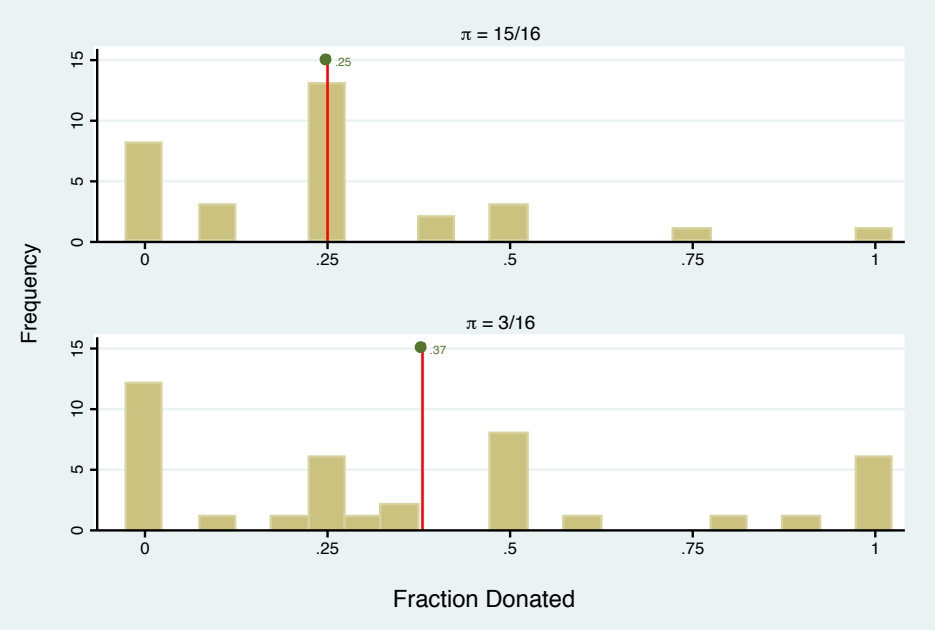
[19] To further minimize participation confounds when subjects see that the mean and median age was 23, we restrict our subjects to the age range of 18 to 30.

[20] Subjects received roughly 30 USD to participate in sessions that run between 30-60 minutes.

numbers between 1 and 16 were randomly chosen to minimize the potential influence of anchoring on the results. They were then asked to write a decision to be placed in a sealed envelope (See Appendix Figure A.2).

Figure 9 reports that participants donated an average amount of 25% when π was high and 38% when π was low. All figures and regression results only analyze the decisions of envelopes opened as we do not have data for envelopes that were shredded. We conducted power calculations to oversample subjects who received low probabilities.[21] In no instance do we impute data from the default decision implemented for a subject when his or her envelope was shredded. For example, when we calculate the expected income of a beneficiary, we use the data for subjects whose envelopes were opened and combine it probabilistically with κ.

FIGURE 9.— Donation and π: Pooled κ



Higher donations when the probability is low may reflect ex-ante fairness concerns. The next figure disaggregates the results by κ. Theory predicts the effect of π to flip depending on the location of κ. Figure 10 shows that the roughly 50% increase in donations is observed in both $\kappa = 0$ and $\kappa = Max$ treatments, which rejects the ex ante fairness explanation for the results in Figure 9.

[21] How many subjects should be assigned to the high and low probability treatments? If we assign the same number of subjects to each, far fewer data will be collected for the treatment condition where only few envelopes are opened. 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 MTurk. This yields a roughly 1:1 ratio for the opened envelopes in the high and low π conditions.

FIGURE 10.— Donation and π: Disaggregated by κ

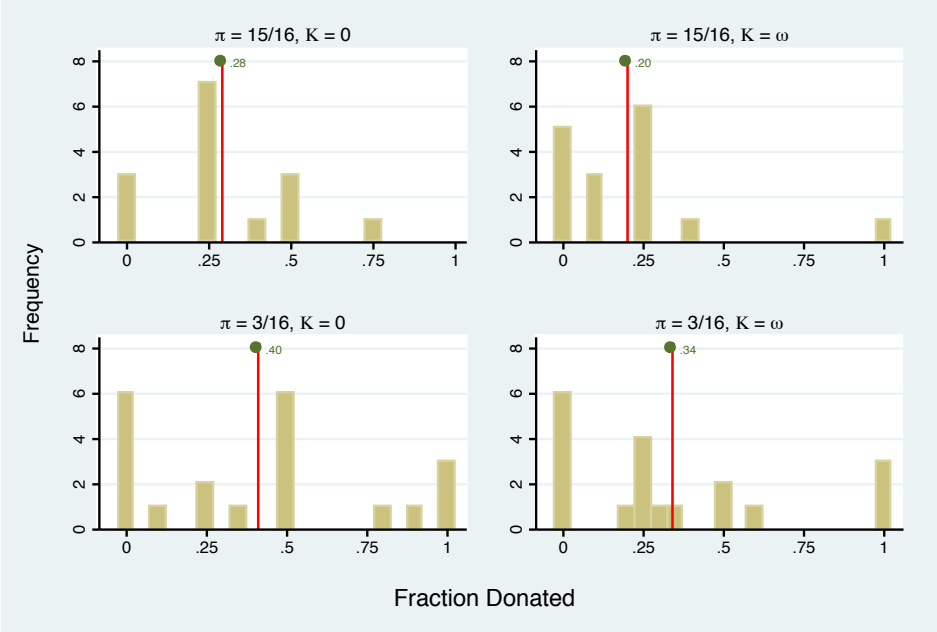


Table IV reports regression results indicating that the change in donations is significant at the 10% level without κ fixed effects (Columns 1) or with κ fixed effects (Column 2). The estimates are stable. The R-square is 0.045, higher than in MTurk, in the regression without any controls. That is, roughly 5% of the variation can be explained by the treatment while the effect size is roughly half the mean donation. Increasing the likelihood of implementation from 0% to 100% reduces the donation by 16-18 percentage points. Columns 3-6 test for ex ante consequentialism. Increasing the likelihood of implementation from 0 to 1 strongly reduces the expected income by the donee by roughly 26% (Columns 3-4) and increases the expected giving of the donor by roughly 22% whether or not κ fixed effects are included (Columns 5-6). These effects are significant at the 1% level.

TABLE IV

DONATION AND π: LINEAR REGRESSION

	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 4 and 5. * p < 0.10, ** p < 0.05, *** p < 0.01

Figure 11 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. When we calculate the expected income of a beneficiary, we use the data for subjects whose envelopes were opened and combine it probabilistically with κ.

FIGURE 11.— Expected Income E(x_2) and π: Disaggregated by κ

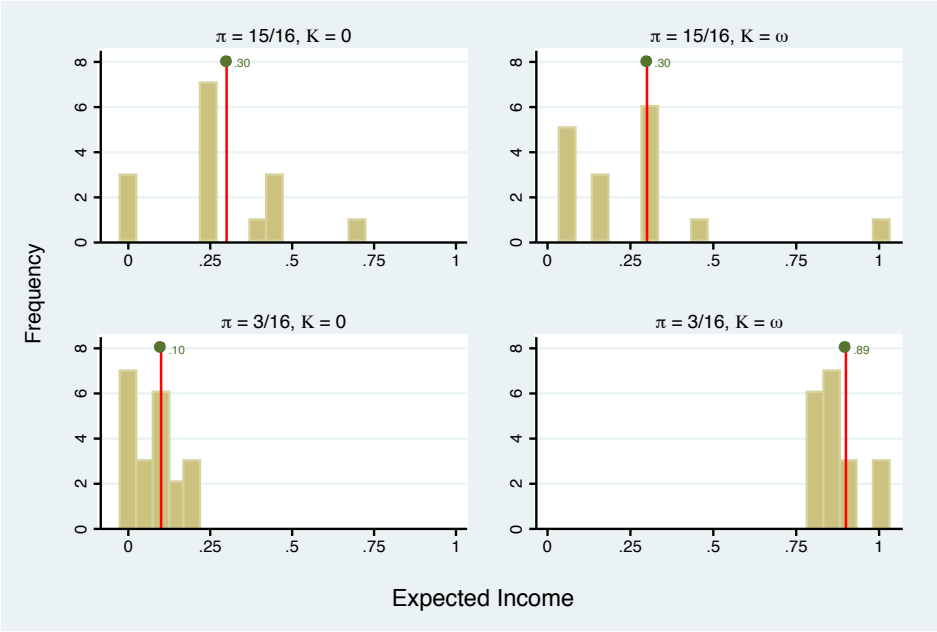
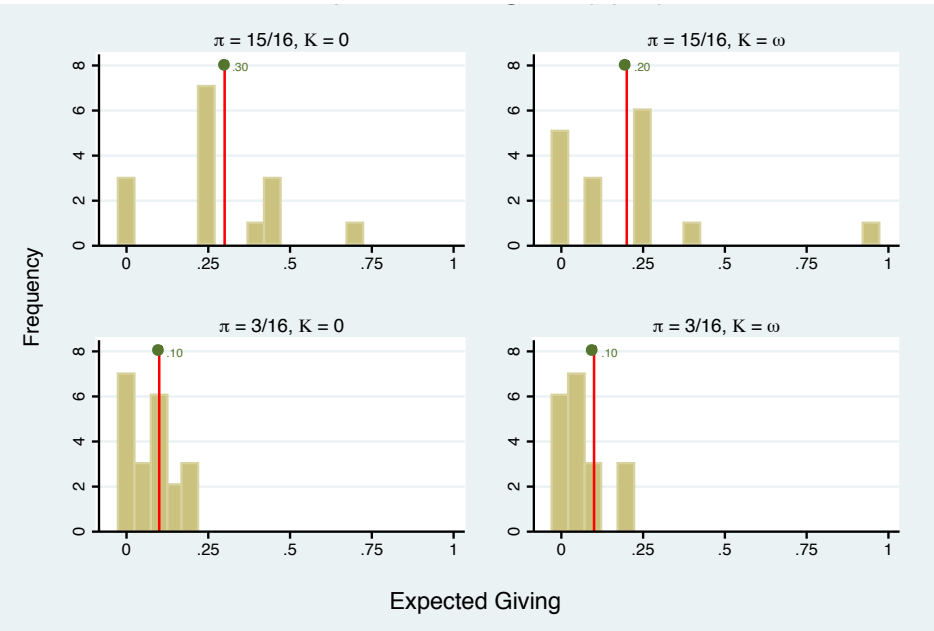


Figure 12 shows that as π changes, expected giving by the decision-maker is also not fixed. Expected giving does not depend on κ. It only depends on d and π. Our results indicate that for both κ, expected giving drops by two-thirds as π goes from high to low. The statistical significance (1% level) of the mean impact is displayed in Columns 5 and 6 of Table IV.

FIGURE 12.— Expected Giving (πd^*) and π: Disaggregated by κ



Finally, we present Mood's median tests in Table V. This tests the null hypothesis that medians of the two populations are identical. It has low power relative to the Mann Whitney test, but is preferred when the variance is not equal in different groups. We can see the variances are different in Figure 9. The median tests report significant differences at the 5% level for π and for κ.

TABLE V

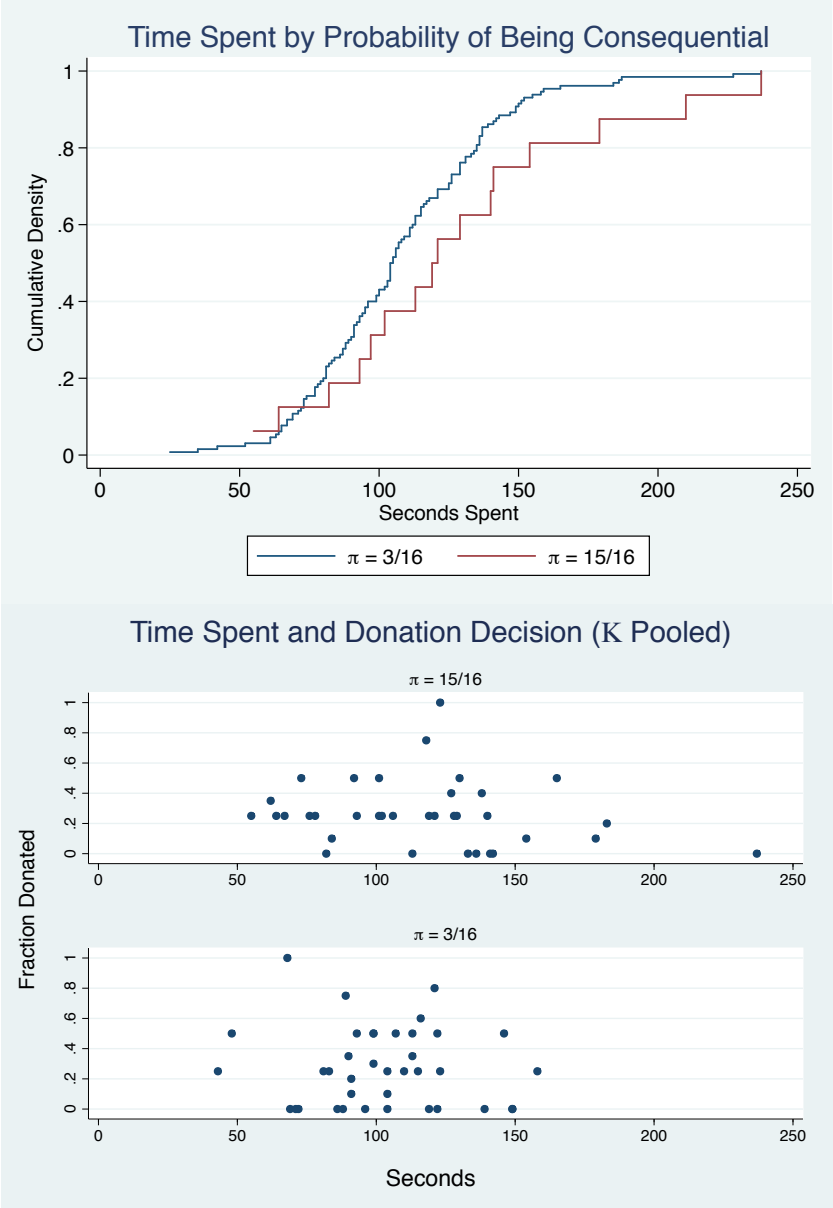
DONATION AND π: NON-PARAMETRIC TESTS

	Non-parametric test for equality of medians, 2-sided test (p-values)
Thresholds	Pooled
$\boldsymbol{\pi} = 3/16$ vs. $\boldsymbol{\pi} = 15/16$	0.04
K = 0 vs. K = Max	0.01

5. COGNITION COSTS

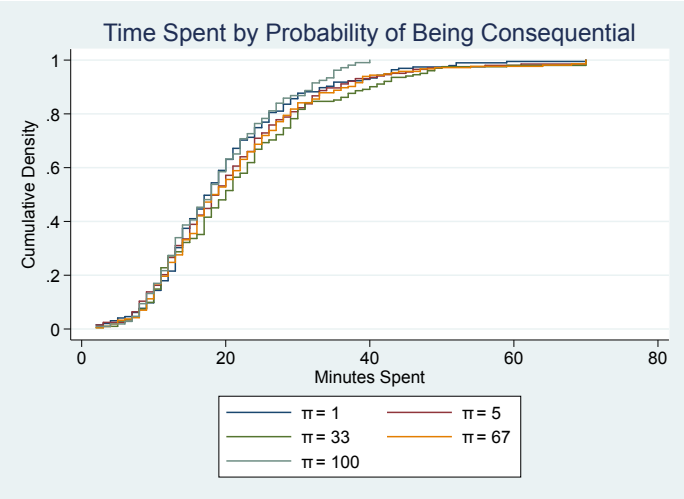
Under the cognitive cost model, individuals spend less time thinking and possibly use altruistic heuristics when their decision is less likely to be implemented. However, individuals spend roughly the same time thinking about their decision regardless of the implementation probability (Figure 13). Moreover, subjects do not donate less when they spend more time on their decision.

FIGURE 13.— Time Spent (on Donation Decision): Lab



On MTurk, we did not have data on the time spent before and after the donation decision and only had data for the entire MTurk session, which is displayed in Figure 14. We find that time spent is only affected (and *reduced*) by $\pi = 1$. This result would appear inconsistent with a cognition costs theory where individuals spend more time on decisions when they are consequential. 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.

FIGURE 14.— Time Spent (Begin vs. End Time): MTurk



One prediction from the cognition cost model is that those whose behaviors are most elastic to π should resort to heuristics more when the probability of being consequential is low. However, Table VI 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 15 shows that those with high $\frac{\delta d}{\delta \pi}$ do not vary their time spent as π changes. The calculations of who are below- or above-median $\frac{\delta d}{\delta \pi}$ are described next.

6. STRUCTURAL ESTIMATION

6.1. *Reduced Form*

To estimate how sensitive the decision d is to π for each individual as predicted from their demographic characteristics, 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:

$$MixedConsequentialistDeontological_i = |\hat{\beta}_0 + \widehat{\beta_1} \mathbf{X}_i|$$

TABLE VI

TIME SPENT (BEGIN VS. END TIME): MTURK HETEROGENEITY BY $\frac{\delta d}{\delta \pi}$

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	

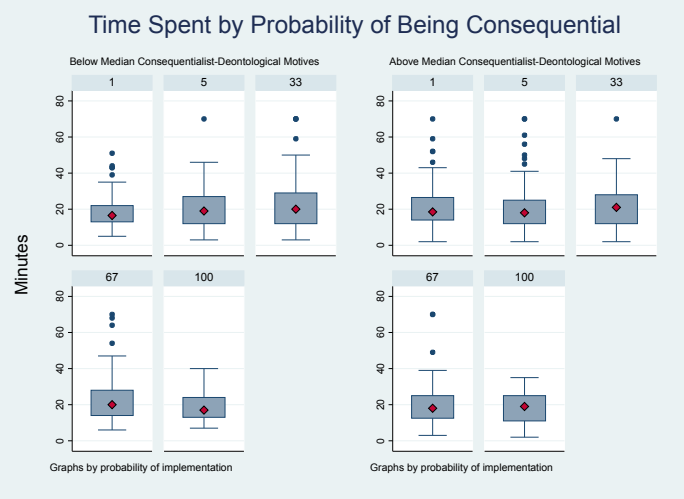
Notes: Standard errors in parentheses. Mixed-Consequentialist aggregates for each subject their demographic characteristics' contribution to the effect of π on the Donation decision. Regressions are weighted by the standard deviation of the first regression to account for uncertainty in the calculation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

We use all the demographic characteristics in $\mathbf{X}_i$ to construct the mixed consequentialist-deontological 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.

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 sizable fractions of people who are pure consequentialist or pure deontological and a large fraction of people who have hybrid motivations. Note that as long as subjects are randomly assignment to treatment, the fact that there are heterogenous types should not bias our results: $\frac{\delta d}{\delta \pi}$ would be constant for pure consequentialists and pure deontologists even if the optimand of their duty is different. We now examine heterogeneity as it correlates with observable demographics.

Table VII 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 $\frac{\delta d}{\delta \pi}$.

FIGURE 15.— Time Spent by $\frac{\delta d}{\delta \pi}$: MTurk



Red Diamond: Median

TABLE VII

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 but still, $\frac{\delta d}{\delta \pi} < 0$. Even when all covariates'

interactions are included, Atheists appear to be the most mixed consequentialist-deontological in their motivations.

To re-emphasize, we do not know if individuals who do not change their behavior are purely deontological or purely consequentialist. We only know that some individuals change their behavior, which we interpret as mixed consequentialist-deontological motivations. We also do not know if a certain share of the population is pure and another share is mixed. We can only infer the presence of at least some people who are mixed. Without additional functional form assumptions, we cannot infer population shares.

6.2. *Estimation*

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.

- 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, 2013b), meaning that subjects view their duty as $d = \delta$ rather than $d \geq \delta$.

- 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.[22]

6.2.1. *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

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

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 I 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%.

6.2.2. *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.[23]

Estimates using two different instruments, being Indian or being under 25, results in similar point estimates (Table VIII). 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 behavior that may appear as consequentialist Fehr-Schmidt preferences may be largely explained by deontological motivations in our experiment and under a bliss point functional form assumption.

7. DISCUSSION

If deontological motivations exist, our model and results suggest an observation of the random

[23]The data also limits our choice of instruments.

TABLE VIII

Donation and π: Linear Regression

	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

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 a few decisions will be implemented. If decision-makers view this decision deontologically, then, decision-makers report more moral decisions than would be the case than if the decision was implemented with certainty, and in experiments that randomize treatment conditions, differences across treatment conditions may be magnified or at least more significant (Bertrand et al. 2004). A similar informal critique bears on policymakers' use of contingent valuation and psychologists' use of vignettes (Diamond and Hausman 1994; List and Gallet 2001; List 2001), though our model shows that any probability in between 100%-consequential and 0%-hypothetical can affect decisions when moral decisions are involved.

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 observation is limited to experiments on social/moral preferences. The random lottery incentive has become the exclusively used incentive system and numerous studies have used and tested it (Hey and Lee 2005). If we a priori conclude that pro-social decisions are invariant to the probability, and if the theorem is correct, then we either reject the existence of deontological motivations[24] or reject FOSD as a behavioral postulate. Stochastic dominance is, however, a compelling criterion for decision-making quality and is generally accepted in decision theory (Quiggin 1990; Wakker 1993; Choi et al. 2014). A broader, formal treatment of the issues for experimental methods along with a literature review, meta-analysis, and new experiments is presented elsewhere (Chen and Schonger 2015).

[24]Or, one could make the unlikely conclusion that invariant agents are purely deontological.

The law also cares about mental states beyond just the consequences: the distinction between *mens rea* (intention) and *actus reus* (act) in criminal law, or opportunistic behavior in equity law, or moral rights in copyright law. Ideological polarization may be related to conflicts of "sacred values" and legal compliance may be driven in part by perceived legitimacy of law, for instance, in the duty that minorities may feel to comply with the law (Tyler and Huo 2002; Chen 2013). The U.S. Department of Defense has begun a research initiative to understand and change sacred values. In general, welfare economics or policy responses that incorporate deontological motivations or estimates of the causal effects of deontological motivations, for example, through the random assignment of judges with different deontological commitments (Chen and Sethi 2011; Chen and Yeh 2014b), is an open question.

Our revealed preference method can distinguish between different internal consequences such as self-image or duty. One response may be that it is a semantic difference. Even purely deontological preferences likely have some neurobiological "consequence" and a future question is what drives deontological motivations, is it conscience or guilt or other parts of the brain? A second response is that shredding removes the experimenter–foreigner presence increased generosity by 20% in dictator games (Cilliers et al. 2015)–so the results can also be interpreted as clean evidence of self-image concerns without experimenter observation. A third response is a distinction between self-image and duty is that the former involves self-signaling as an investment that has long-term consequences (Bénabou and Tirole 2011) so self-image and duty can be distinguished through forgetting: an agent who remembers the good deed may be motivated by self-image but an agent that forgets (and knows that she will forget) may be motivated by duty. For example, consider an individual doing a good deed knowing that he resides in Plato's cave.[25] The Mechanical Turk experiment, where the decision is less salient, is more likely to involve forgetting.[26]

8. 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 suggests that individuals also seem to care about decisions *per se*, independently of their consequences. Deontological motivations are frequently cited in law and philosophy. Moral vignettes have been used by experimental philosophers and psychologists to identify deontological motivations. This paper asks the behavioral question: Do people have deontological motivations? Answering this question is challenging because real-world behavior that may seemingly be motivated deontologically may actually be motivated for consequentialist reasons. This paper formalizes the notion of consequentialist as well as deontological motivations as properties of preference relations; suggests and implements a thought experiment inspired by Kant's categor-

[25] What we mean is something analogous to an individual knowing that he will take the blue pill (cf. the 1999 film The Matrix) and forget everything.

[26] In contrast, a public randomization device with shredding is more memorable.

ical imperative that uses revealed preference to detect deontological motivations; and outlines the relevance of deontological motivations for economics and law.

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. We hope that our study will contribute in generating further work on revealed preference estimation methods that could be more efficient in detecting deontological motivations, with or without the presence of additional motivations.

REFERENCES

Alexander, Larry, and Michael Moore, 2012, *Stanford Encyclopedia of Philosophy*.

Alger, Ingela, and Jörgen W Weibull, 2013, Homo Moralis-Preference Evolution Under Incomplete Information and Assortative Matching, *Econometrica* 81, 2269–2302.

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

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

Anscombe, Francis J, and Robert J Aumann, 1963, A definition of subjective probability, *Annals of mathematical statistics* 199–205.

Arrow, Kenneth Joseph, 2012, *Social Choice and Individual Values*, Cowles Foundation Monographs Series, third edition (Yale university press, New Haven), Monograph 12.

Batson, C. Daniel, Judy G. Batson, Jacqueline K. Slingsby, Kevin L. Harrell, Heli M. Peekna, and R. Matthew Todd, 1991, Empathic Joy and the Empathy-Altruism Hypothesis, *Journal of Personality and Social Psychology* 61, 413–426.

Battigalli, Pierpaolo, and Martin Dufwenberg, 2007, Guilt in Games, *The American Economic Review* 97, 170–176.

Bénabou, Roland, and Jean Tirole, 2006, Incentives and Prosocial Behavior, *The American Economic Review* 96, 1652–1678.

Bénabou, Roland, and Jean Tirole, 2011, Identity, Morals, and Taboos: Beliefs as Assets, *The Quarterly Journal of Economics* 126, 805–855.

Bentham, Jeremy, 1791, *Panopticon* (T. Payne, London).

Berdejó, Carlos, and Daniel L. Chen, 2014, Priming Ideology? Electoral Cycles without Electoral Incentives among U.S. Judges, Working paper, ETH Zurich.

Bergstrom, Theodore C., Rodney J. Garratt, and Damien Sheehan-Connor, 2009, One Chance in a Million: Altruism and the Bone Marrow Registry, *The American Economic Review* 99, 1309–1334.

Bertrand, Marianne, Esther Duflo, Sendhil Mullainathan, et al., 2004, How Much Should We Trust Differences-in-Differences Estimates?, *The Quarterly Journal of Economics* 119, 249–275.

Besley, Timothy, 2005, Political selection, *The Journal of Economic Perspectives* 19, 43–60.

Binmore, Ken, 1994, Playing fair: Game theory and the social contract, *Cambridge, Mass.: MIT Press* .

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

Brock, J. Michelle, Andreas Lange, and Erkut Y. Ozbay, 2013, Dictating the Risk: Experimental Evidence on Giving in Risky Environments, *The American Economic Review* 103, 415–437.

Camerer, Colin, 2011, The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List, *Available at SSRN 1977749* .

Cappelen, Alexander W, Trond Halvorsen, Erik Sorensen, and Bertil Tungodden, 2013a, Face-saving or fair-minded: What motivates moral behavior?, *NHH Dept. of Economics Discussion Paper* .

Cappelen, Alexander W., Astri Drange Hole, Erik Ø Sørensen, and Bertil Tungodden, 2007, The Pluralism of Fairness Ideals: An Experimental Approach, *American Economic Review* 97, 818–827.

Cappelen, Alexander W, James Konow, Erik Ø Sørensen, and Bertil Tungodden, 2013b, Just luck: An experimental study of risk-taking and fairness, *The American Economic Review* 103, 1398–1413.

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

Chen, Daniel L., 2006, Islamic Resurgence and Social Violence During the Indonesian Financial Crisis, in Mark Gradstein, and Kai A. Konrad, eds., *Institutions and Norms in Economic Development*, chapter 8, 179–199 (MIT Press).

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

Chen, Daniel L., 2011, Markets and Morality: How Does Competition Affect Moral Judgment?, Working paper, Duke Law School.

Chen, Daniel L., 2013, The Deterrent Effect of the Death Penalty? Evidence from British Commutations During World War I, Working paper, ETH Zurich.

Chen, Daniel L., 2014a, Can Markets Overcome Repugnance? Muslim Trade Reponse to Anti-Muhammad Cartoons, Working paper, ETH Zurich, Mimeo.

Chen, Daniel L., 2014b, Can Markets Stimulate Rights? On the Alienability of Legal Claims, *RAND Journal of Economics* 46, 23–65.

Chen, Daniel L., and John J. Horton, 2014, Are Online Labor Markets Spot Markets for Tasks?: A Field Experiment on the Behavioral Response to Wage Cuts, *Management Information Systems Quarterly* (Revised and Resubmitted).

Chen, Daniel L., Vardges Levonyan, S. Eric Reinhart, and Glen B. Taksler, 2014, Mandatory Disclosure: Theory and Evidence from Industry-Physician Relationships, Working paper, Mimeo.

Chen, Daniel L., and Jo Lind, 2007, Religion, Welfare Politics, and Church-State Separation, *Journal of Ecumenical Studies* 42, 42–52.

Chen, Daniel L., and Jo Thori Lind, 2014, The Political Economy of Beliefs: Why Fiscal and Social Conservatives and Liberals Come Hand-in-Hand, Working paper.

Chen, Daniel L., Moti Michaeli, and Daniel Spiro, 2015a, Ideological Perfectionism on Judicial Panels, Working paper, ETH Zurich.

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue, 2015b, Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires, *The Quarterly Journal of Economics* Revise and Resubmit.

Chen, Daniel L., and Martin Schonger, 2015, A Theory of Experiments: Invariance of Equilibrium to the Strategy Method of Elicitation and Implications for Social Preferences, Working paper, ETH Zurich.

Chen, Daniel L., and Jasmin Sethi, 2011, Insiders and Outsiders: Does Forbidding Sexual Harassment Exacerbate Gender Inequality?, Working paper, University of Chicago.

Chen, Daniel L., and Susan Yeh, 2013, How Do Rights Revolutions Occur? Free Speech and the First Amendment, Working paper, ETH Zurich, Mimeo, Zurich.

Chen, Daniel L., and Susan Yeh, 2014a, The Construction of Morals, *Journal of Economic Behavior and Organization* 104, 84–105.

Chen, Daniel L., and Susan Yeh, 2014b, Growth Under the Shadow of Expropriation? The Economic Impacts of Eminent Domain, Working paper, ETH Zurich and George Mason University.

Choi, Hyunkyung, Marcia Van Riper, and Suzanne Thoyre, 2012, Decision making following a prenatal diagnosis of Down syndrome: an integrative review, *Journal of Midwifery & Womens Health* 57, 156–164.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman, 2014, Who is (More) Rational?, *The American Economic Review* 104, 1518–1550.

Cilliers, Jacobus, Oeindrila Dube, and Bilal Siddiqi, 2015, The White-Man Effect: How Foreigner Presence Affects Behavior in Experiments, *Journal of Economic Behavior and Organization* .

Dana, Jason, Daylian M Cain, Robyn M Dawes, et al., 2006, What you don't know won't hurt me: Costly (but quiet) exit in dictator games, *Organizational Behavior and Human Decision Processes* 100, 193–201.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang, 2007, Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness, *Economic Theory* 33, 67–80, Symposium on Behavioral Game Theory.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier, 2013, Voting to Tell Others, Working paper.

Diamond, Peter A., and Jerry A. Hausman, 1994, Contingent Valuation: Is Some Number better than No Number?, *The Journal of Economic Perspectives* 8, 45–64.

Elizabeth Hoffman, Matthew L. Spitzer, 1985, Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice, *The Journal of Legal Studies* 14, 259–297.

Ellingsen, Tore, and Magnus Johannesson, 2008, Pride and Prejudice: The Human Side of Incentive Theory, *The American Economic Review* 98, 990–1008.

Falk, Armin, and Urs Fischbacher, 2006, A Theory of Reciprocity, *Games and Economic Behavior* 54, 293–315.

Falk, Armin, and Nora Szech, 2013, Morals and Markets, *Science* 340, 707–711.

Feddersen, Timothy, Sean Gailmard, and Alvaro Sandroni, 2009, Moral Bias in Large Elections: Theory and Experimental Evidence, *The American Political Science Review* 103, 175–192.

Fehr, Ernst, and Klaus M. Schmidt, 1999, A Theory of Fairness, Competition, and Cooperation, *The Quarterly Journal of Economics* 114, 817–868.

Fischbacher, Urs, 2007, z-Tree: Zurich toolbox for ready-made economic experiments, *Experimental economics* 10, 171–178.

Friedman, Milton, and Leonard J. Savage, 1948, The Utility Analysis of Choices Involving Risk, *The Journal of Political Economy* 56, 279–304.

Gibson, Rajna, Carmen Tanner, and Alexander F. Wagner, 2013, Preferences for Truthfulness: Heterogeneity among and within Individuals, *The American Economic Review* 103, 532–548.

Gneezy, Uri, 2005, Deception: The Role of Consequences, *The American Economic Review* 95, 384–394.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen, 2001, An fMRI Investigation of Emotional Engagement in Moral Judgment, *Science* 293, 2105–2108.

Grossman, Zachary, 2015, Self-signaling and social-signaling in giving, *Journal of Economic Behavior & Organization* 117, 26–39.

Harsanyi, John C, 1977, Morality and the theory of rational behavior, *Social Research* 623–656.

Hey, John D, and Jinkwon Lee, 2005, Do subjects separate (or are they sophisticated)?, *Experimental Economics* 8, 233–265.

Holt, Charles A., 1986, Preference Reversals and the Independence Axiom, *The American Economic Review* 76,

508–515.
Kant, Immanuel, 1797, Über ein vermeintes Recht aus Menschenliebe zu lügen, *Berlinische Blätter* 1, 301–314.
Kant, Immanuel, 1959, *Foundations of the Metaphysics of Morals*, II edition (Prentice Hall), trans. by L. W. Beck.
Kaplow, Louis, and Steven Shavell, 2001, Any Non-welfarist Method of Policy Assessment Violates the Pareto Principle, *Journal of Political Economy* 109, 281–286.
Kaplow, Louis, and Steven Shavell, 2006, *Fairness Versus Welfare* (Harvard University Press).
Konow, James, 2000, Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions, *The American Economic Review* 90, 1072–1091.
Kreps, David M, 1988, *Notes on the Theory of Choice* (Westview Press Boulder).
Levhari, David, Jacob Paroush, and Bezalel Peleg, 1975, Efficiency Analysis for Multivariate Distributions, *The Review of Economic Studies* 42, 87–91.
List, John 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, John A., and Craig A. Gallet, 2001, What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?, *Environmental and Resource Economics* 20, 241–254.
Machina, Mark J., 1982, "Expected Utility" Analysis without the Independence Axiom, *Econometrica* 50, 277–323.
Machina, Mark J., 1989, Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty, *Journal of Economic Literature* 27, 1622–1668.
Machina, Mark J, and David Schmeidler, 1992, A more robust definition of subjective probability, *Econometrica: Journal of the Econometric Society* 745–780.
Mankiw, N. Gregory, and Matthew Weinzierl, 2010, The Optimal Taxation of Height: A Case Study of Utilitarian Income Redistribution, *American Economic Journal: Economic Policy* 2, 155–176.
McCabe, Kevin A., Mary L. Rigdon, and Vernon L. Smith, 2003, Positive reciprocity and intentions in trust games, *Journal of Economic Behavior & Organization* 52, 267–275.
Mikhail, John, 2007, Universal Moral Grammar: Theory, Evidence and the Future, *Trends in Cognitive Sciences* 11, 143 – 152.
Nagel, Thomas, 1970, *The Possibility of Altruism* (Clarendon Press, Oxford).
Nisbett, Richard E, and Dov Cohen, 1996, *Culture of Honor: The Psychology of Violence in the South* (Westview Press).
Nozick, Robert, 1974, *Anarchy, State, and Utopia*, Harper Torchbooks (Basic Books).
Orne, Martin 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, John, 1982, A Theory of Anticipated Utility, *Journal of Economic Behavior & Organization* 3, 323–343.
Quiggin, John, 1990, Stochastic Dominance in Regret Theory, *The Review of Economic Studies* 57, 503–511.
Rabin, Matthew, 1993, Incorporating Fairness into Game Theory and Economics, *The American Economic Review* 83, 1281–1302.
Rand, David G., Joshua D. Greene, and Martin A. Nowak, 2012, Spontaneous Giving and Calculated Greed, *Nature* 489, 427–430.
Riker, William H., and Peter C. Ordeshook, 1968, A Theory of the Calculus of Voting, *The American Political Science Review* 62, 25–42.
Roth, Alvin E., 2007, Repugnance as a Constraint on Markets, *The Journal of Economic Perspectives* 21, 37–58.
Savage, Leonard J, 1972, *The Foundations of Statistics* (Courier Corporation).
Shayo, Moses, and Alon Harel, 2012, Non-consequentialist voting, *Journal of Economic Behavior & Organization* 81, 299–313.
Sinnott-Armstrong, Walter, 2012, Consequentialism, in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*.
Smith, Adam, 1761, *The Theory of Moral Sentiments* (A. Millar).
Smith, Kyle D., John P. Keating, and Ezra 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.
Smith, Vernon L, 2016, The Fair and Impartial Spectator .
Starmer, Chris, 2000, Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk, *Journal of Economic Literature* 38, 332–382.
Starmer, Chris, and Robert Sugden, 1991, Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation, *The American Economic Review* 81, 971–978.
Titchener, James L., 1967, Experimenter Effects in Behavioral Research, *Archives of Internal Medicine* 120, 753–755.
Tversky, Amos, and Daniel Kahneman, 1992, Advances in Prospect Theory: Cumulative Representation of Uncertainty, *Journal of Risk and Uncertainty* 5, 297–323.
Tyler, Tom R, 1997, The psychology of legitimacy: A relational perspective on voluntary deference to authorities, *Personality and social psychology review* 1, 323–345.

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

Wakker, Peter, 1993, Savage's Axioms Usually Imply Violation of Strict Stochastic Dominance, *The Review of Economic Studies* 60, 487–493.

Wilcox, Nathaniel 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 $\kappa = 0$. Subjects come to this page after finishing three IQ tasks.

FIGURE A.1.— Screenshot in Lab Experiment

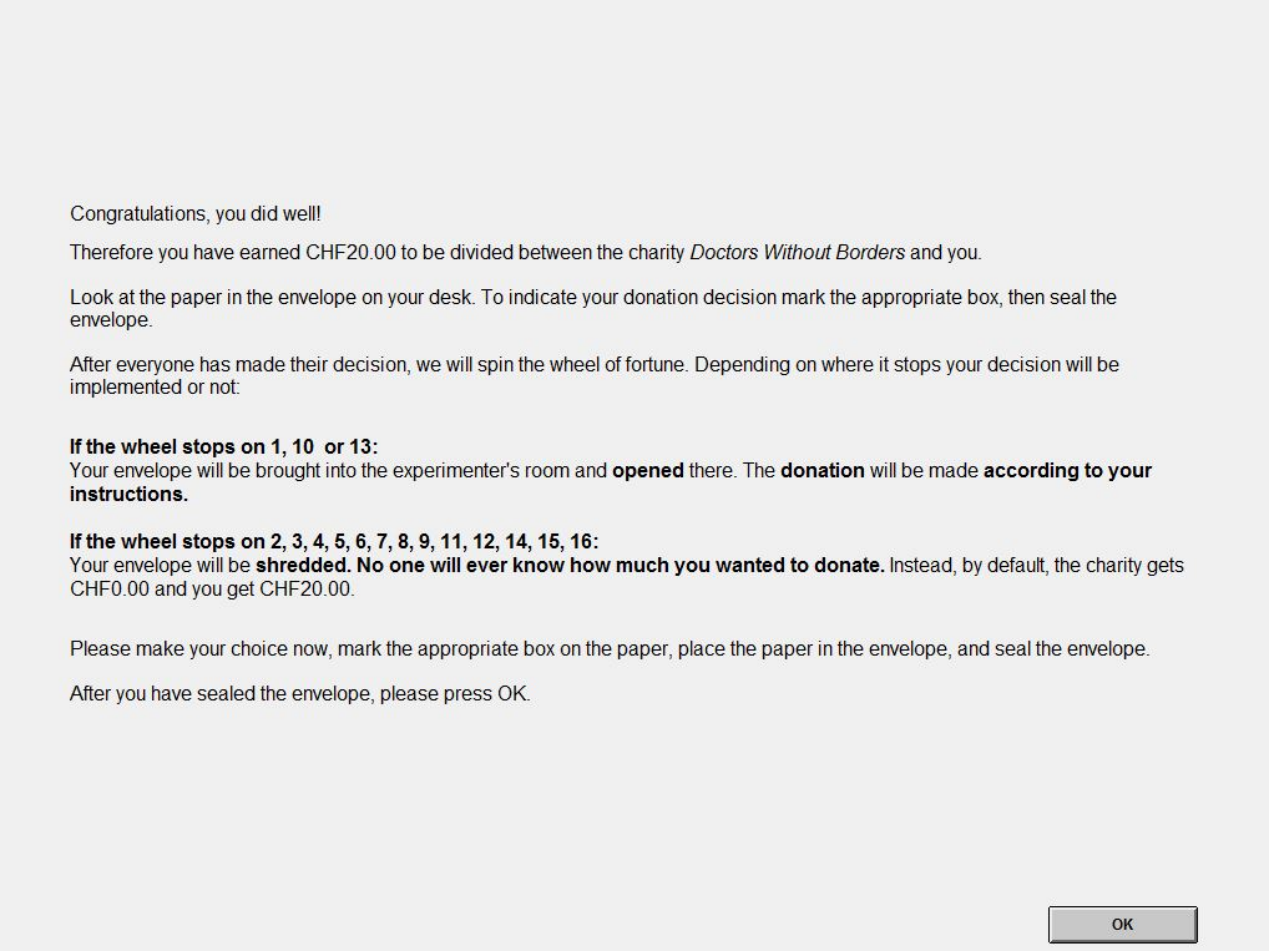


FIGURE A.2.— Donation Decision Placed in Sealed Envelope

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!

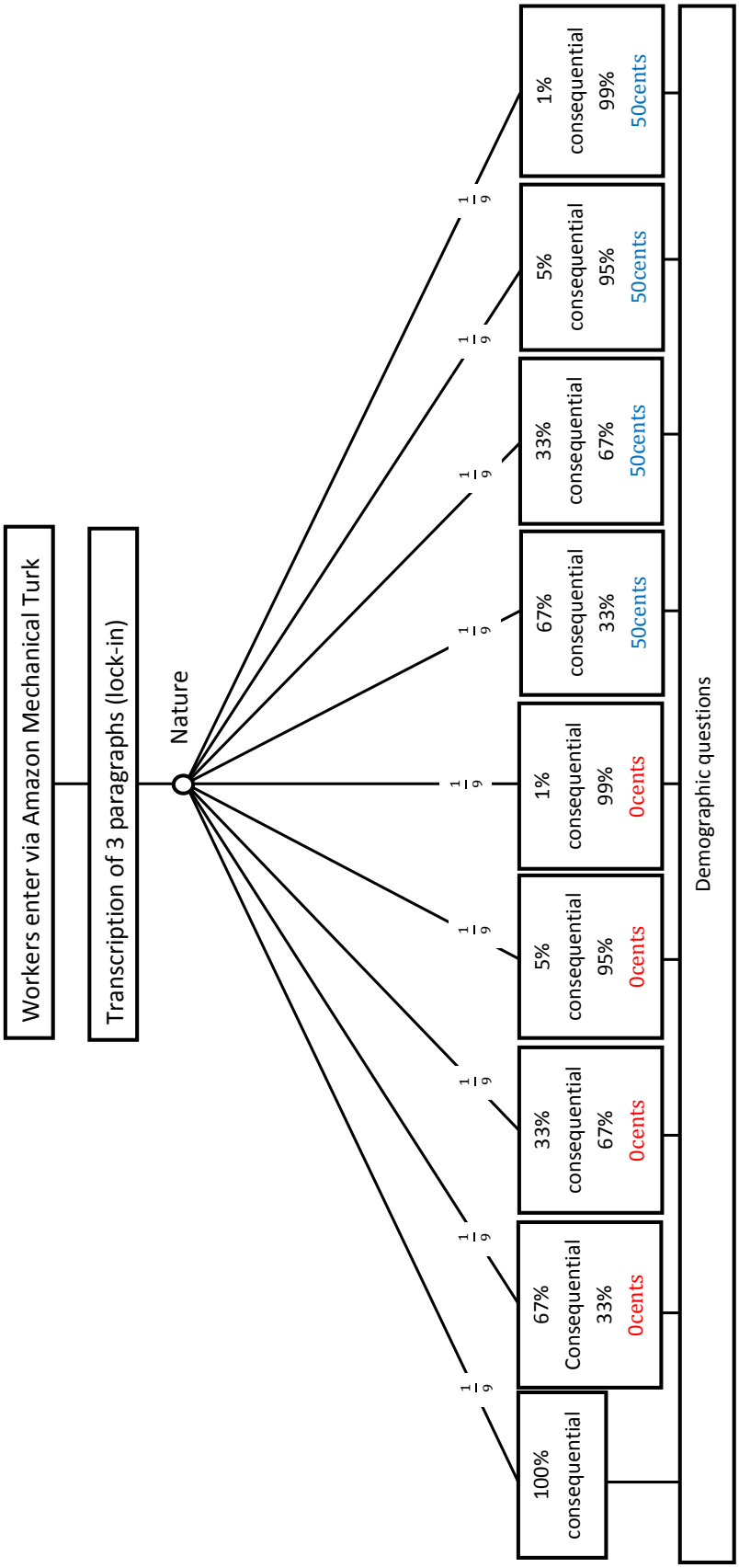


FIGURE A.3.— Schematic of MTurk Experiment (Experiment 1)

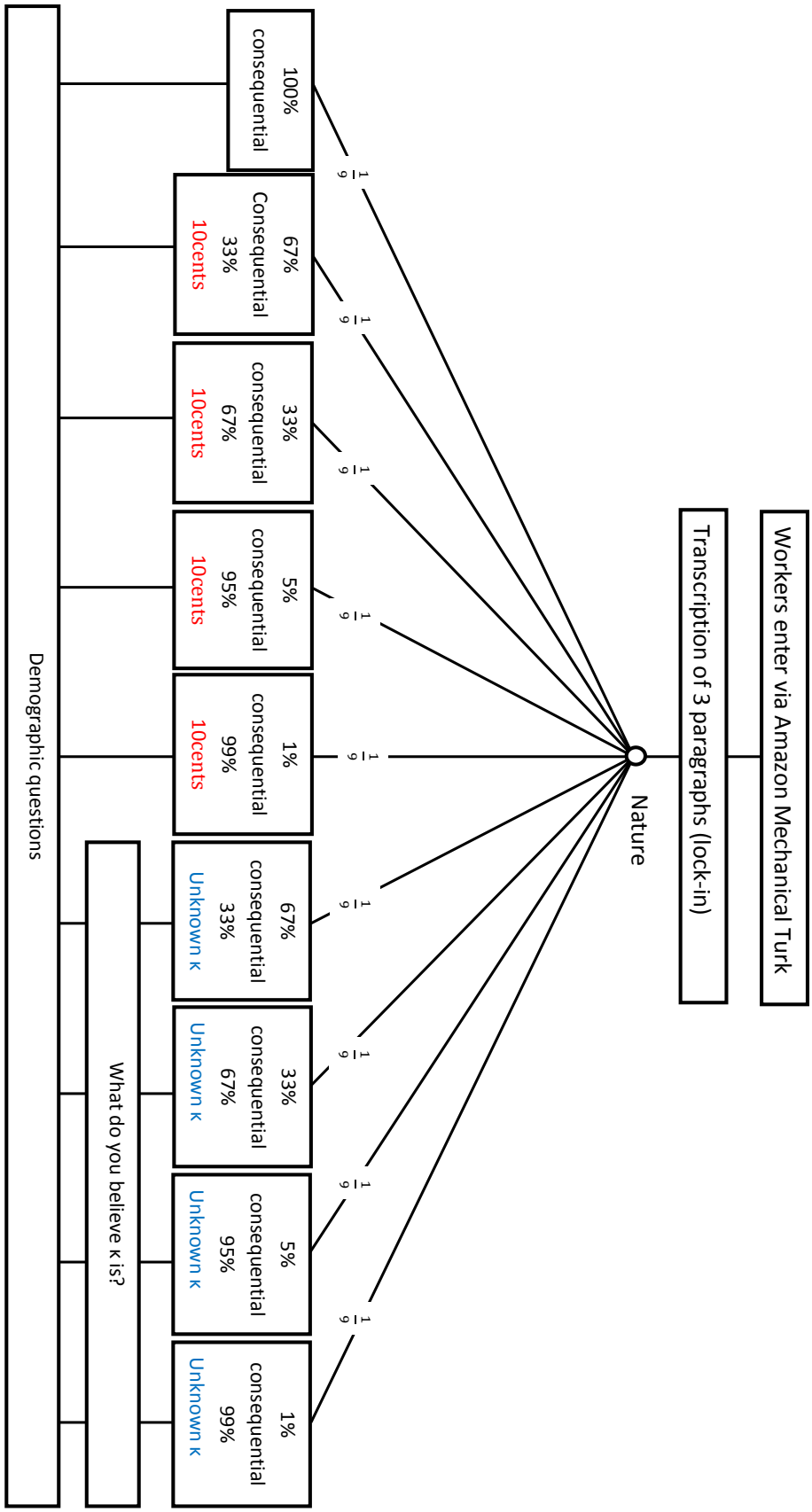


FIGURE A.4.— Schematic of MTurk Experiment (Experiment 2)