SOCIAL SCIENCES

Social preferences or sacred values? Theory and evidence of deontological motivations

Daniel L. Chen[1]* and Martin Schonger[2]

Recent advances in economic theory, largely motivated by experimental findings, have led to the adoption of models of human behavior where decision-makers take into consideration not only their own payoff but also others' payoffs and any potential consequences of these payoffs. Investigations of deontological motivations, where decision-makers make their choice based on not only the consequences of a decision but also the decision per se, have been rare. We provide a formal interpretation of major moral philosophies and a revealed preference method to distinguish the presence of deontological motivations from a purely consequentialist decision-maker whose preferences satisfy first-order stochastic dominance.

> Your friend is hiding in your house from a murderer. The murderer arrives and asks you whether your friend is hiding in your house. Assuming you cannot stay silent, should you lie or tell the truth? (*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 noneconomists, that policy should be determined deontologically—people, society, and judges have duties; from duties, they derive what is the correct law, right, and just. This paper asks the behavioral question: Are there deontological motivations? If so, how would these motivations be formally modeled? What do deontological motivations imply for economics? What puzzles can be explained that elude standard models?

In the past few decades, economic theory has gradually expanded the domain of preferences. The homo oeconomicus view that individuals are only motivated by self-regarding material consequences confronted mounting evidence, usually in the laboratory, that individuals had other motivations—such as fairness [e.g., (*2*)], inequality aversion [e.g., (*3*)], reputation [e.g., (*4*–*7*)], or social image [e.g., (*8*, *9*)]. A common feature of these models is that motivations are consequentialist, in the sense that preferences are over acts because of their effects. These preferences are prominently characterized as hypothetical imperatives—preferences over acts because of their consequences—as opposed to categorical imperatives—preferences over acts regardless of their consequences—which Kant (*1*) called deontological motivations.

In general, the presence of deontological motivations is hard to detect. The usual method to measure deontological motivations is through survey or vignettes that present ethical dilemmas, like the moral trolley problem (*10*). What our paper develops is a revealed preference method and a theorem that predicts invariance in the thought experiment if people are motivated solely under consequentialist motivations; however, if deontological motivations are present, in combination with consequentialist ones, then this thought experiment will reveal variance.

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, warm glow, 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 nonconsequential state, d remains unknown to anyone, including the experimenter. With consequentialism, preferences are over lotteries (*11*). With deontological motivations, d matters per se, even in the nonconsequential state. To illustrate, Kant said in his axe-murderer hypothetical, "You must not lie," no matter the consequences.

Think of d^1, d^2, …, $d^{|D|}$, as possible decisions. Our experiment varies the probability that the decision is implemented. With some probability, π, your decision is implemented—has consequences—and with probability $1-\pi$, your decision has no consequence. Thus, x^C is a function of the decision, and x^N is some constant outcome that is invariant to your decision. This thought experiment can apply to any decision with a moral element, but we illustrate our theorem using the dictator game as it is one of the games most used in the academic literature. In a dictator game, you have your endowment ω, and you can donate anywhere from 0 to ω. In our thought experiment, 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. Subjects put their irrevocable decisions anonymously in sealed envelopes, and their envelope is shredded with some probability with a public randomization device and the probability is known in advance (Fig. 1). Shredding means that the decision has no consequences, not even through the experimenter, by eliminating motivations related to experimenter observation (*12*) and any altruism related to the societal good of providing one's data for science. The decision only has consequences if the envelope is opened. Our shredding criterion for deontological motivations parallels Kant's discussion of his own thought experiment. Kant, likewise, allowed for uncertainty—the possibility that the decision has the ultimate adverse consequence or has no consequences—but "to be truthful in all declarations is a sacred and unconditionally commanding law of reason that admits no expediency whatsoever." Kant's categorical imperative focused on the act itself rather than the expected consequences of an act. It is this motivation that we seek to model and uncover behaviorally.

The closest field analogs of our experiment may be found in two recent papers. First, Bergstrom *et al.* (*13*) examined the decision to

[1]Toulouse School of Economics, Institute for Advanced Study in Toulouse, University of Toulouse Capitole, Toulouse, France. [2]Center for Law and Economics, ETH Zürich, Zurich, Switzerland.
*Corresponding author. Email: daniel.chen@iast.fr

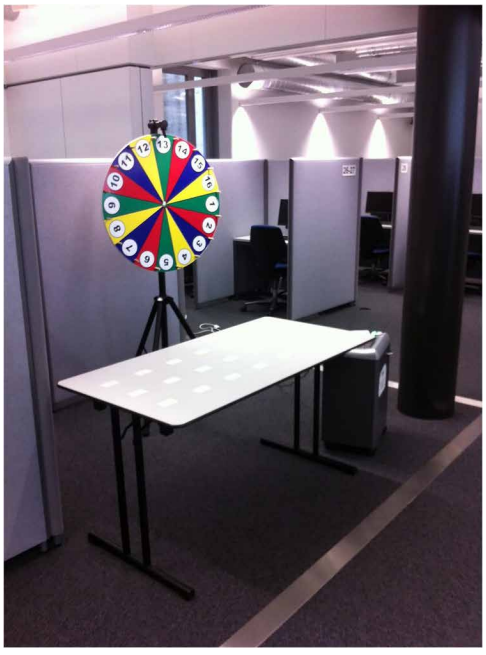

Fig. 1. Laboratory implementation. Subjects put their irrevocable decisions anonymously in sealed envelopes, and their envelope is shredded with some probability with a public randomization device. Photo credit: Martin Schonger, ETH Zürich.

sign up as a bone marrow donor. With some probability, the decision to sign up has consequences, such that the recipient receives bone marrow and the donor undergoes expensive and painful surgery. Bergstrom *et al.* (*13*) found that those less likely to sign up to be a bone marrow donor came from ethnic groups that, due to genetic match and need, were more likely to be called off the list to donate. They argue this pattern to be a puzzle. Second, Choi *et al.* (*14*) studied the decision not to abort a fetus with Down syndrome. Prospective parents varied in the probability that the decision to abort would have consequences. They found that as the prospect became more real (hypothetical, high risk, versus diagnosed), parents were more likely to abort. In both (*13*) and (*14*), as π decreased, people became more likely to choose a decision that might be interpreted as deontological. However, in both settings, d is not irrevocable and not anonymous and π is not exogenous, leaving room for potential confounders. In our laboratory setting, d is irrevocable and anonymous and π is exogenously assigned to the individual.

Formally, we show that pure deontologists following the categorical imperative would not change their behavior as the probability changes, but, counterintuitively, it turns out that pure consequentialists also do not change their behavior. We provide a graphical and formal proof that someone who satisfies the behavioral assumption of first-order stochastic dominance (FOSD) and is purely consequentialist will not change their behavior as the probability changes. Simply put, the DM is choosing between lotteries G and F, so if G first-order stochastically dominates F with respect to $\succsim$ [i.e., if for all x': $\sum_{x:x' \succsim x} G(x) \leq \sum_{x:x' \succsim x} F(x)$] and then if a decision d is optimal for one probability π, then it is the optimal d for all probabilities. As a corollary, we can state the result with expected utility (a stronger behavioral assumption than FOSD). For the DM donating the marginal penny, the marginal benefit of donating is the recipient's well-being and any social consequence of that increase. The marginal cost is to give up that penny. The DM equates the marginal benefits and marginal costs. As the probability that the decision is implemented falls, then both the marginal benefits and costs fall equally, so the DM still makes the same decision on the margin because the indirect objective function is proportional to the utility of the decision implemented with certainty.

To bridge our theorem to experimental evidence, our first study uses subjects in a laboratory. We asked subjects to choose an amount for a charitable recipient (as illustrated in Fig. 2), a third-party aid organization. We found that subjects became 50% more charitable as the decision becomes more hypothetical. Our second piece of evidence uses an online anonymous experiment, allowing large samples and very low implementation probabilities; but a difference is that d is observed by the experimenter even in the nonconsequential state. If motives related to the experimenter or the study are strong, then we may expect less variance. We found that subjects became 33% more charitable as the decision becomes more hypothetical.

It is possible that subjects become more charitable as the implementation probability falls because they value some kind of ex ante fairness involving preferences over expected outcomes (*15*, *16*, *17*). While this is not a deontological motivation in Kant's typology, it is a behavioral motivation that can confound the interpretation of our results. To investigate that motive, the two experiments also had a treatment arm where the nonconsequential state involves the entire sum being donated. Our data can rule out an expected-income targeter, who should have become less generous in response to reductions in π. Our data can also rule out other ex ante fairness motivations. Last, our data on decision time suggest that cognition costs are also not the explanation for variance between high and low π.

Our third piece of evidence illustrates how assumptions on the curvature of motives together with data on decision variance can inform how individuals trade-off between consequentialist and deontological motives. We use standard parameterizations of a structural model—consequentialist motivations are estimated with a classic Fehr-Schmidt inequity aversion utility, while deontological motivations are estimated as a bliss point as in (*18*, *19*). The variation in our data generated by the experiment is consistent with largely deontological rather than consequentialist motives under the entire range of standard inequity aversion parameters.

Like Bergstrom *et al.* (*13*) observing more bone marrow donations and Choi *et al.* (*14*) observing more decisions to not abort when the decisions were more hypothetical, we see d increases when π falls. What our model suggests is that as the probability falls, the (net negative) consequences of carrying out the act falls, but the (deontological) benefits of the act remain high. Moreover, the direction of change can give insight into the location of the maximand for an individual's duty (relative to the consequentialist maximand). Assuming the pure deontologist's maximand is higher than the pure consequentialist's maximand, reducing the probability results in decisions that are more deontological.

Our paper makes two contributions to the economic literature—theoretical and experimental. Economic models have thus far focused on hypothetical imperatives (preferences over acts because of their consequences). This interpretation is supported by Sobel's (*20*) extensive literature review of interdependent preferences, part of which offered a typology of non–homo oeconomicus models. In one class are Chicago School models that model preferences over general commodities transformed into consumption goods. In another class are identity models [e.g., (*21*)] with utility functions over

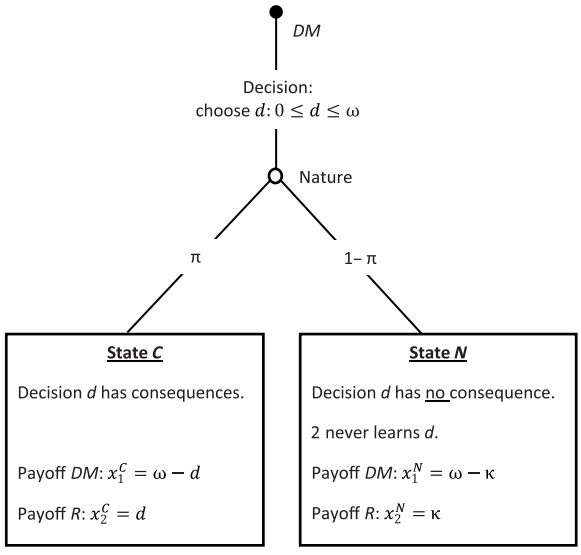

Fig. 2. Schematic of the experiment. An irrevocable decision is implemented with a probability.

actions and an identity that incorporates the prescriptions that indicate the identity-appropriate behavior. Sobel noted that "the models of Akerlof–Kranton and Stigler–Becker are … mathematically identical. It is curious that these formally equivalent approaches are associated with schools of thought that often are viewed as opposites. The theories are identical because they are consistent with precisely the same set of observations." In our reading, both classes of models fall under the hypothetical imperative: Chicago agents choose between quantities but do not have preferences over choices versus preferences over quantities. In identity models, agents choose acts but do not have preferences over acts versus preferences over consequences of acts. The categorical imperative would distinguish these preferences. Our thought experiment and shredding criterion likewise distinguishes choices from quantities and acts from consequences of acts.

Empirical researchers also have assumed that choices do not enter the utility function separate from the causal effects of choices. For example, in the random lottery incentive, experimental subjects make many choices, but only one of them is chosen at random to be implemented. In this oft-used method in experimental economics, when decisions involve a deontological element, the degree of pro-social behavior may be over-estimated. The lower the likelihood of implementation in the random lottery incentive, the greater the over-estimation of pro-social behavior. In the strategy method—another method often used to increase statistical power—subjects make many choices corresponding to possible states that may depend on what other subjects choose but only a fraction of decisions count for pay. Deontological motives would imply that this bias from random lottery incentives would never disappear, no matter how high the stakes are.

Likewise, in surveys (which includes contingent valuation), subjects report preferences in nonconsequentialist settings (e.g., valuation of an environmental good in a hypothetical scenario), and the decisions may change as the decision becomes more likely to be implemented. In measuring willingness to pay, subjects report a price that is implemented if it is higher than a randomly generated price in the Becker-DeGroot-Marschak method. In the Vickrey auction, bidders submit written bids that are consequential only for the highest bidder. The higher the price, the more likely the decision has consequences. In market design data, subjects report preferences over choices over schools whose likelihood of being consequential varies.

Notably, our operationalization of deontological motives—choosing a decision regardless of the likelihood of implementation (i.e., irrespective of the consequences)—bears close similarity to the concept of legitimacy defined in psychology. Tyler (*22*) considered laws and organizations to be legitimate if these laws and organizations motivate obedience to rules irrespective of likelihood of reward or punishment. The remainder of the paper is organized as follows. The related literature is presented next. Then, in Results, we define consequentialism, deontologicalism, and mixed motivations; we prove that behavior is invariant to the probability for pure consequentialism and for pure deontologicalism, but varies for mixed motivations. Subsequently the empirical evidence is described. We conclude with a discussion, and a description of materials and methods.

Related literature

Smith's (*23*) impartial spectator in *The Theory of Moral Sentiments* may have been deontological though perhaps also consequentialist.

> "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." (*23*).

There is a vast economics literature on concepts related to deontological motivations. We refer the reader to Sobel's (*20*) extensive literature review and focus our discussion here to subsequent work.

The three closest theoretical developments may be as follows. First, deontological motivations may relate to identity investment. In (*24*), moral decision-making is modeled as a form of identity investment that prevents future deviant behavior. Here, motives can be deontological or consequentialist. The DM cares about the fact that the decision is implemented. Second, deontological motivations may also relate to expressive motives. People may participate in elections even when their vote is not pivotal because of a perceived duty to vote (*25*). Feddersen *et al.* (*26*) and Shayo and Harel (*27*) formalize the insight where individuals obtain a small positive payoff by the act of voting for an option independent of the electoral outcome, which they test with experiments by varying the probability of being pivotal. Here, expressive motives can be deontological or consequentialist. The DM cares about the fact that the vote is cast. Election outcomes are public, so a message is sent to the public and vote share can affect the legitimacy of a candidate. DellaVigna *et al.* (*28*) show experimentally that the act of voting includes motives to tell others. Third, deontological motivations may also relate to "homo kantiensis," whose preferences are ones that are socially optimal when everyone else also holds that view (*29*). Alger and Weibull (*29*) report that these 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. Here, motives can be deontological or consequentialist. The DM cares about the outcome of everyone making the same decision.

Warm glow motives can also be deontological or consequentialist. In an earlier theoretical contribution, Andreoni (*30*) 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, the author suggests in (*9*) that social audience motivations can provide microfoundation for the warm glow. Thus, the DM cares about the fact that the decision is observed. In other work, Ellingsen and Johannesson (*31*) have a utility function incorporating the DM's payoff, others' payoff, and how others think of the DM. The DM cares about the consequences of actions. Deontological motivations may also relate to guilt aversion (*32*). The prototypical cause would be the infliction of harm or distress on the recipient, which can be deontological or consequentialist.

A large experimental literature has been interested in studying the motives for prosocial behavior. The shredding criterion can be distinguished from the experimental paradigm that varies the probability that one's decision will have an impact, because in those paradigms, the DM experiences the cost of helping in both states of the world (*33*, *34*). In other experimental paradigms (*26*, *27*, *35*, *36*), the DM experiences the benefits of the decision in both states of the world. In a contemporaneous research design that is related, Andreoni and Bernheim (*9*) use a modified dictator game with random implementation probabilities, but there are five differences. First, we make the recipient a charitable organization outside the laboratory; in their study, the recipients are in the room observing the decision and dictators become more generous as the probability of implementation increases because they are motivated by their social audience. Second, we make both the probability and the realization of the state of nature public; in their study, recipients observe the probability but not the fact that nature chose the outcome. Third, in their study, they acknowledge that there may be motivations regarding what the experimenter infers and regard this as a confound; our laboratory experiment shreds decisions, which directly removes that confound. Fourth, their study uses the strategy method and subjects play several games, whereas in our study, each subject sees only one probability and we do not use the strategy method. Fifth, they recognize the importance of not using within-subject variation for any particular game; we directly remove sequence effects and contrast effects (for example, if an experimenter asks two questions with a higher and lower probability, then subjects may feel that the right answer is to give more in one scenario, which would be a confound for our invariance theorem). In another contemporaneous study, Grossman (*35*) 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. The study does not shred the decisions, so the decision's contribution is still a consequence. More broadly, we rule out motives related to the beliefs of others because the third-party aid organization is unaware of the subject.

Large literatures outside of economics, such as psychology, political science, sociology, and law, have discussed concepts related to deontological motives. Sacred values and taboos are also often interpreted as pertaining to duty, and some actions cannot be evaluated through costs and benefits (*37*). Some of these have been analyzed by economists—conflicts of sacred values (*38*), repugnance (*39*, *40*), and saving the lives of mice (*41*). Besley (*42*) has argued to screen for deontological motivations in business leaders, politicians, or judges. In contrast, Kaplow and Shavell (*43*) criticize relying on nonconsequentialist motivations in optimal policy design as it would necessarily harm some individuals.

RESULTS

In *The Stanford Encyclopedia of Philosophy*, Sinnott-Armstrong (*44*) defines 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 (*45*); any welfarist view is consequentialist (*46*). 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." (*47*).

We introduce our thought experiment and focus on this definition of 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.

Thought experiment

The idea to identify nonconsequentialist motivations by varying the probability of the DM's decision being consequential guides this paper. The DM has a real-valued choice variable d that influences both her own monetary payoff x_1 and 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 $\pi > 0$, and state N occurs 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; 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 3 illustrates this.

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

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

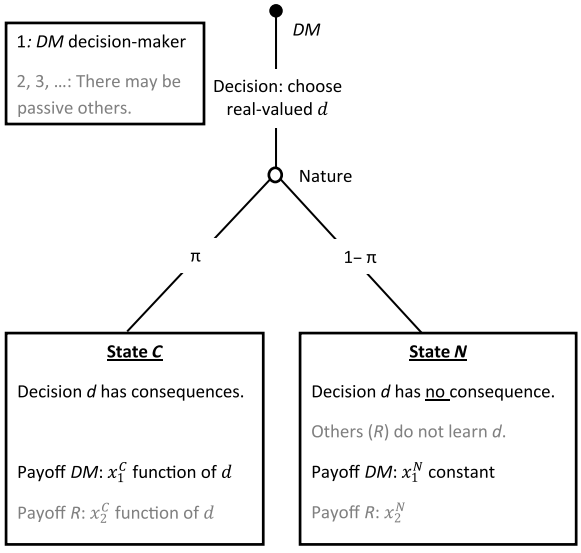

Fig. 3. Schematic of the thought experiment. The process for making a charitable decision.

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

Limiting attention to pure consequentialists, the problem simplifies to

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

and the indirect objective function to

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

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

Graphical proof

In the previous subsection, we have seen that if the DM satisfies the axioms of expected utility and if d^* is not constant in the probability, then she cannot be a consequentialist. Put differently, if we observe a DM to vary her decision in the probability, then we would reject the joint hypothesis that the DM is a consequentialist and an expected-utility maximizer. Because expected utility theory often fails to describe behavior (*48*) 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 FOSD 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 of FOSD and is purely consequentialist will not change their behavior as the probability changes. The left-hand side of Fig. 4 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 because instead of getting 3, sometimes, the DM gets 4. Formally, G first-order stochastically dominates F with respect to $\gtrsim$ if for all x': $\sum_{x:x' \gtrsim x} G(x) \leq \sum_{x:x' \gtrsim x} F(x)$.

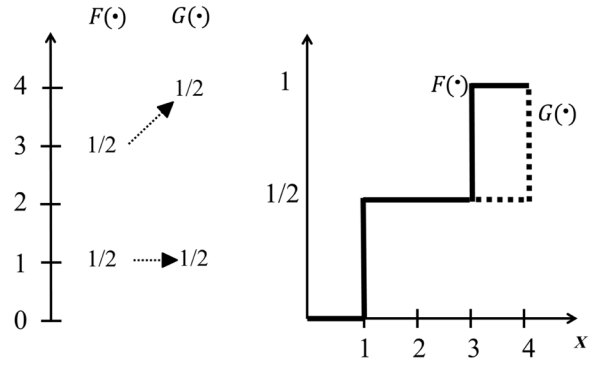

Fig. 4. First-order stochastic dominance. A textbook example of FOSD.

For every outcome x', the probability of any outcome worse than x' is lower under G than under F. The right-hand side of Fig. 4 provides an example of such cumulative distribution functions (CDFs). For the proof, recall that decisions are choices over lotteries like F and G. Suppose 1 is the nonconsequentialist outcome, and let 3 or 4 be the active choice. What does changing the probability do? It moves the horizontal bar up and down. However, G always FOSD F. Hence, if a choice is optimal for one probability, then it is the optimal choice for all probabilities.

Formal statement of assumptions and theorem

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. It may seem odd to model deontological motivations by utility functions because one may view "utility" as a consequence, but because ours is a revealed preference approach, we follow the usual economics approach (*49*) of modeling DMs' behavior as if they maximized that objective function and refrain from interpreting the function as standing for utility or happiness.

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. In the general case with all motivations present, the Bernoulli utility function satisfies $u = u(x_1, x_2, d)$. The standard theories of decision-making by Savage (*50*) and Anscombe and Aumann (*11*) rely on the assumption that the domain of consequences is state-independent.

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 that they are simply mirroring the other extreme of consequentialist preferences and could thus be represented by $u = u(d)$. However, because 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. A deontologist can be formalized as having a lexicographic preference on decisions d and outcome x, with deontological before consequentialist motivations.

Definition 3. Deontological preferences: A preference is called deontological if there exist u and 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 $f(x) \geqq f(x')]$.

It is possible to model purely deontological people as having a different choice set (*51*). However, 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 DM as internal constraints, that is, as the first part of preferences in a lexicographic framework. The reason that 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 that 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) (*47*).

We delineate assumptions that allows us to experimentally identify with observable choice behavior whether subjects have preferences where both motivations are present (i.e., whether their preferences belong to the category of consequentialist-deontological preferences). The standard consequentialist approach to (and a central assumption for) choice under uncertainty is FOSD. A wide variety of models of choice under uncertainty satisfies FOSD and thus falls within this framework; among them, most prominent are not only the expected utility theory and its generalization by Machina (*52*) but also the cumulative prospect theory (*53*) or rank-dependent utility theory (*54*).

Following the canonical framework as laid out by Kreps (*55*), 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 DM.

Axiom. (Preference order) Let $\succsim$ be a complete and transitive preference on P.

This is the standard axiom saying that the preference relation is a complete ordering. It implicitly includes consequentialism because the preference relation is on P, that is, over lotteries that are over consequences x.

Next we define FOSD. Often, definitions of FOSD are suitable only for preference orders that are monotonic in the real numbers [for example, see (*56*)]. 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.)

Definition. (FOSD) p first-order stochastically dominates q with respect to the ordering induced by $\succsim$, if for all x': $\sum_{x:x' \succsim x} p(x) \leq \sum_{x:x' \succsim x} q(x)$.

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

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

Formally, our theorem needs both strict FOSD and weak FOSD because strict FOSD does not imply weak FOSD.

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.

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

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. Proofs and additional theoretical discussion are relegated to section S1.

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, and 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 $u(d)$.

Consequentialist-deontological preferences

Next, we illustrate consequentialist-deontological preferences where the optimal decision changes as the probability of being consequentialist changes. For exposition, we do so in the context of Fig. 2 and simplify notation such that the net consequences are a function of x_1.

Example 1. $u = u(x_1, d) = x_1 + b(d)$, 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 $b_{11}(d) < 0$. 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 $\frac{\partial d^*}{\partial \pi} = 0$.

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 $\frac{\partial d^*}{\partial \pi} = \frac{f_1(\omega - d^*)}{\pi f_{11}(\omega - d^*) + b_{11}(d^*)} < 0$, because utility is increasing in its own outcomes and the denominator that 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).

For more complicated utility functions, nonadditive or nonglobally 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

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

Thus, 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.

Potential confounds

Ex ante fairness

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). In our experimental setup, for example, a subject might target the expected income of the recipient and thus vary the decision in the probability. This section shows formally that by varying κ, we can test whether people have these ex ante considerations. 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: one where κ is set at 0 and another where κ is set at the maximum.

Example 2. Targeting the recipient's expected income. Consider the following preferences: $U(x_1, x_2) = E[x_1] + a(E[x_2]) = \pi x_1^C + (1 - \pi)x_1^N + a(\pi x_2^C + (1 - \pi)x_2^N)$. 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: $U = f(E[u(x_1)], E[\tilde{u}(x_2)])$, 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), $u, \tilde{u}$ is $u_1, \tilde{u}_1 > 0$ (strictly increasing), $u_{11}, \tilde{u}_{11} \leq 0$ (weakly concave), and $\pi > 0$. Then, the indirect objective function is

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

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

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

Hence, there exists a unique solution. The first-order condition for this problem is $\frac{\tilde{u}_1(d)}{u_1(\omega - d)} - \frac{f_1}{f_2} = 0 \equiv F$. The FOC (first-order condition) 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 $\frac{\partial^2 V(d)}{(\partial d)^2} < 0$: $\mathrm{sgn}\left(\frac{\partial d^*}{\partial \pi}\right) = \mathrm{sgn}\left(\frac{\partial F(d^*,\pi)}{\partial \pi}\right)$. It can be shown that

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

Thus, 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 $\frac{\partial d^*}{\partial \pi} \leq 0$; while, for $\kappa = \omega$, $\frac{\partial d^*}{\partial \pi} \geq 0$.

Thus, experimentally, by varying κ, we can test whether people have these ex ante considerations. In summary, 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.

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 framework because they are incurred during the decision and are a consequence that even arises if the nonconsequential 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 DM solve the metaproblem optimally, see (*57*). This section shows that a cognition-costs model would predict that (i) time spent on the survey also changes with π as d changes. Our research design also provides a second test: (ii) Subjects with greater cognition costs should have $\frac{\delta d}{\delta \pi} = 0$ for a larger range of π near 0.

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. Suppose the heuristic choice tends to be a cooperative or fair one (*58*), 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 nonheuristic choice, $V(d) = \pi u(\omega - d, d, \gamma) + (1 - \pi)u(\omega - \kappa, \kappa, \gamma)$. Define $\check{d} \equiv \mathrm{argmax} V(d)$. Obviously, $\check{d}$ does not vary in π. The DM will choose to act heuristically if $V(\check{d}) < V(\bar{d})$ or

$$F(\pi) \equiv V(\check{d}) - V(\bar{d}) = \pi(u(\omega - \check{d}, \check{d}, \gamma) - u(\omega - \bar{d}, \bar{d}, 0))$$

$$+(1-\pi)(u(\omega - \kappa, \kappa, \gamma) - u(\omega - \kappa, \kappa, 0)) < 0$$

Because $(1-\pi)(u(\omega-\kappa,\kappa,\gamma) - u(\omega-\kappa,\kappa,0)) < 0$, 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 $\tilde{\pi}$ with $0 < \tilde{\pi} < 1$ such that $F(\tilde{\pi}) = 0$, 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 0, the heuristic is better. Because $\check{d} \neq \bar{d}$, this means that these 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 that we vary the cognition cost, that is, we do an exercise in comparative statics and investigate how $\tilde{\pi}$ varies in γ, and note that

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

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, $\tilde{\pi}$ is a constant function of γ, but there, the above assumptions are violated. Figure 5 shows when, as a function of a probability, someone would incur a given cognition cost. Hence, if we could experimentally vary not only probability but also cognition costs and then observe it, then 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 DMs switching to a heuristic $\bar{d}$, which may be higher or lower than the preferred choice $\check{d}$, leading to the inability to infer consequentialist-deontological preferences. If DMs 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 (i) time spent on the survey also changes with π as d changes. We also provide a second test: (ii) 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 the cognition costs actually incurred results. The higher the cognition cost parameter, the further to the right and the larger the S-shape. Figure 5 illustrates this, plotting the cognition cost incurred (γ) against the probability of being consequential (π) for two cognition cost parameters, γ^L and γ^H, where $\gamma^L < \gamma^H$. The dotted line is for the subject experiencing low cognition costs, while the dashed line is for the subject experiencing high cognition costs.

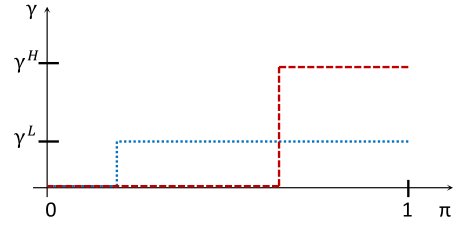

Fig. 5. S-shape cognition costs. The cognition costs as thinking harder about a decision creates cognition costs.

Self-image

A conceptual distinction can be made between self-image and duty. First, in economic models of self-image motives, decisions are affected when subjects anticipate finding out about peers (*59*). Because self-image is related to ego, individuals may punish those who threaten their ego. In addition, self-image is often modeled as an investment with long-term consequences (*24*). These motives depart from the Kantian duty described earlier.

Laboratory experiment

Participants donated an average amount of 25% when π was high and 38% when π was low. Figure 6 disaggregates the results by κ, and the vertical lines indicate means for each treatment group. Ex ante fairness concerns would predict the effect of π to flip depending on the location of κ, but we observed an increase in donations (of roughly 50%) for both $\kappa = 0$ and $\kappa = \text{Max}$ treatments.

Table 1 reports regression results, indicating that the change in donations is significant at the 10% level without κ fixed effects (column 1) or with κ fixed effects (column 2). The estimates are stable. The R^2 is 0.045 only including π. The magnitude of the effect is equivalent to roughly half the mean donation. Extrapolating linearly suggests that increasing the likelihood of implementation from 0 to 100% reduces the donation by roughly 17 percentage points. Columns 3 to 6 test for ex ante consequentialism. Increasing the likelihood of implementation from 0 to 1 strongly reduces the expected income by the donee (columns 3 and 4) and strongly increases the expected giving of the donor (columns 5 and 6), whether or not κ fixed effects are included. These effects are significant at the 1% level. The following presents additional visualizations of these results.

Figure 7 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 8 shows that as π changes, expected giving by the DM 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 1.

Table 2 presents Mood's median tests of 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 that the variances are different in Fig. 6. The median tests report significant differences at the 5% level for π and for κ.

Online experiment

Figure 9 shows that the lower the π, the more generous is the DM. The increase in generosity is monotonic with the decrease in probability. Donations increased from 18% (when $\pi = 1$) to 27% (when $\pi = 0.01$). The following presents regression results, and we can again strongly reject the hypothesis that subjects are targeting expected income or expected giving.

Table 3 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%. Column 2 adds demographic controls. Country of origin was coded as United States and

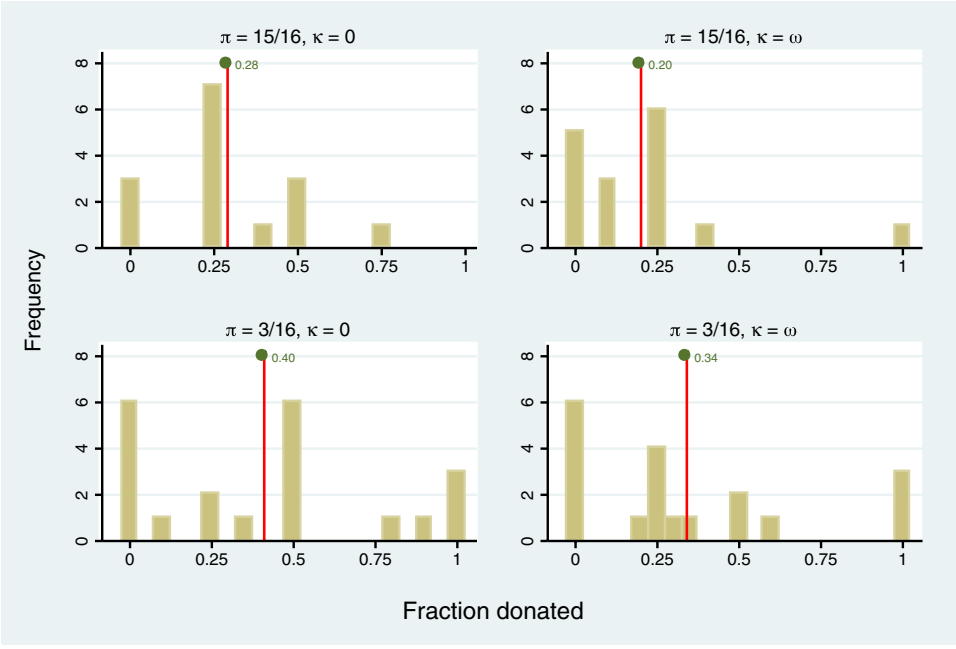

Fig. 6. Donation and π: Disaggregated by κ. Donation data from the laboratory experiment. The vertical lines indicate the mean donation of each treatment group.

Table 1. Donation and π: Linear regression. This table presents regression results from the laboratory experiment. SEs in parentheses. Raw data shown in Figs. 6 and 7. *$P < 0.10$, **$P < 0.05$, and ***$P < 0.01$.

	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)
κ 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

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. Columns 3 and 4 consider if subjects target expected income, and columns 5 and 6 consider 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%. To make calculations on expected donations when κ is unknown, we use data on perceived donation.

Table 4 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. The effects are not significantly different across treatment arms.

We next examine whether the distributions of donation decisions are significantly affected by π. Table 5 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 π = 1,0.67, and 0.33 is the same as the distribution of decisions for subjects treated with π = 0.05 and 0.01. The lower panel of Table 5 reports that the distribution of donations does not significantly vary by κ. Means are also not significantly different by κ.

Next, we reject cognition costs as the driving feature for decision change. The three findings are as follows: (i) individuals spend roughly the same time thinking about their decision regardless of the implementation probability, (ii) donations were not associated with time spent, and (iii) those estimated to be most responsive to

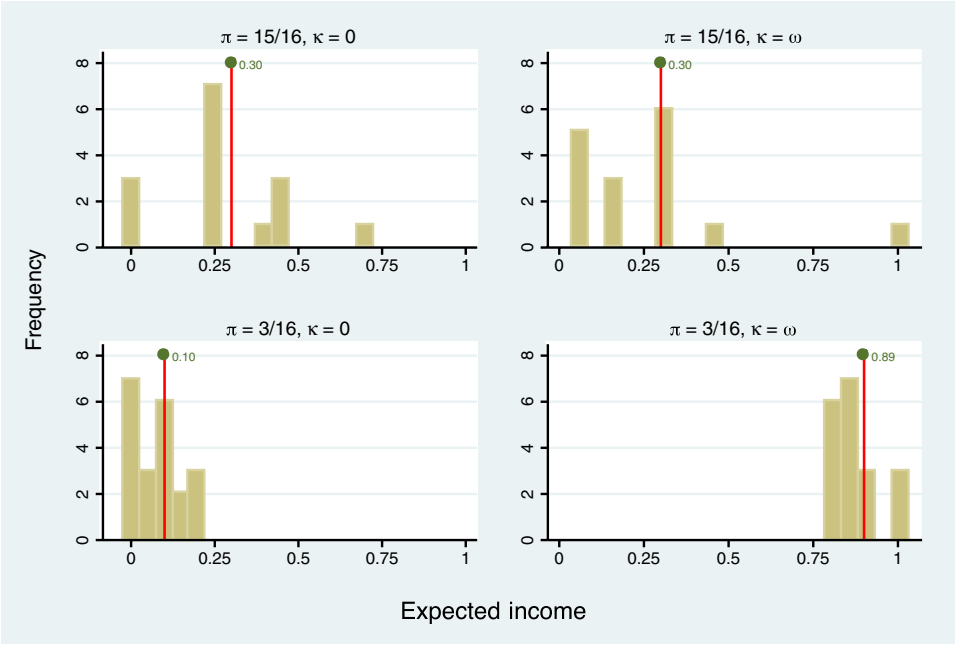

Fig. 7. Expected income E(x_2) and π: Disaggregated by κ. The expected income using the decision data from the laboratory experiment. The vertical lines indicate the mean expected income of each treatment group.

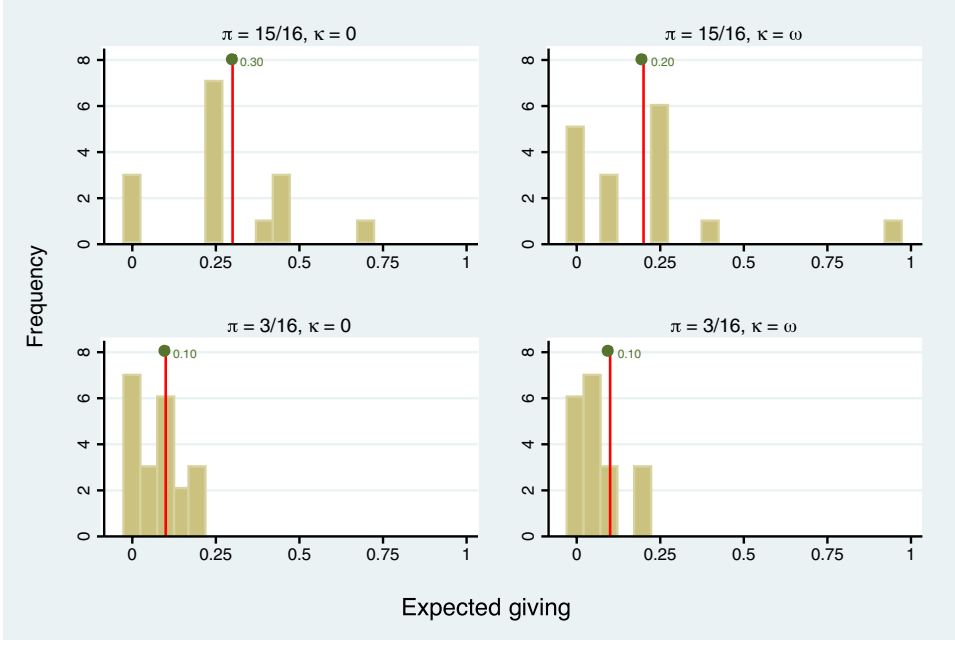

Fig. 8. Expected giving (πd^*) and π: Disaggregated by κ. The expected giving using the decision data from the laboratory experiment. The vertical lines indicate the mean expected giving of each treatment group.

implementation probability do not seem to be resorting to heuristics more, at least measured by time spent.

Figure 10 shows that individuals spend roughly the same time thinking about their decision regardless of the implementation probability, which is inconsistent with the cognitive cost model, where individuals spend less time thinking and use altruistic heuristics when their decision is less likely to be implemented. Moreover, subjects do not donate less when they spend more time on their decision to compensate for cognition effort.

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 Fig. 11. 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.

Table 6 shows that, at low π, those with below-median $\frac{\delta d}{\delta \pi}$ spend less time than those with above-median $\frac{\delta d}{\delta \pi}$ (see below for an explanation for how these groups are determined). In addition, Fig. 12 shows that those with high $\frac{\delta d}{\delta \pi}$ do not vary their time spent as π changes. These findings are inconsistent with the cognition cost model in that those whose behaviors are most elastic to π (high $\frac{\delta d}{\delta \pi}$) do not seem to be resorting to heuristics more when the probability of being consequential is low, at least measured by time spent.

Table 2. Donation and π: Nonparametric tests. This table presents Mood's median tests of the null hypothesis that medians of the two populations are identical.

	Nonparametric test for equality of medians, two-sided test (*P* values)
Thresholds	Pooled
$\pi = 3/16$ versus $\pi = 15/16$	0.04
$K = 0$ versus $K = $ Max	0.01

Table 7 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 $\frac{\delta d}{\delta \pi}$.

Structural estimation

This section presents structural estimates of how individuals trade-off between consequentialist and deontological motivations. We provide two illustrations. First, we follow Cappelen *et al.* (*18*, *19*) and assume that homogenous individuals maximize homo oeconomicus consequentialist motivations but place weight λ on a deontological portion that follows bliss point preferences: $u(x_{DM}, x_2, d) = \lambda(x_1) + (-(\delta - d)^2) = \lambda(1 - d) + (-(\delta - d)^2)$ (Note that this means that the model by Cappellen *et al.* views duty as $d = \delta$ rather than $d \geq \delta$. We assume that subjects' duties are enumerated in percent terms). The first-order condition is $0 = \pi\lambda(-1) + 2(\delta - d)$, which 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, representing the decision when $\pi = 0$. Figure 9 would yield a bliss point $\delta = 0.25$, which is very close to the observed 27% when $\pi = 0.01$. Then, because we can pin down one of two unknown parameters, we can identify the weight placed on deontological motivations using the speed of change as π varies; in this case, $\lambda = 0.14$. Note that a pure homo oeconomicus would maximize d^* at 0, which is why λ increases monotonically with speed of change.

Our second illustration models consequentialist motivations as in Fehr and Schmidt (*3*), plugging in α and β inequality parameters for $u(x_{DM}, x_2, d) = \lambda(x_1 - \alpha\max\{x_2 - x_1, 0\} - \beta\max\{x_1 - x_2, 0\}) + (-(\delta - d)^2)$. The individual's first-order condition over their choice

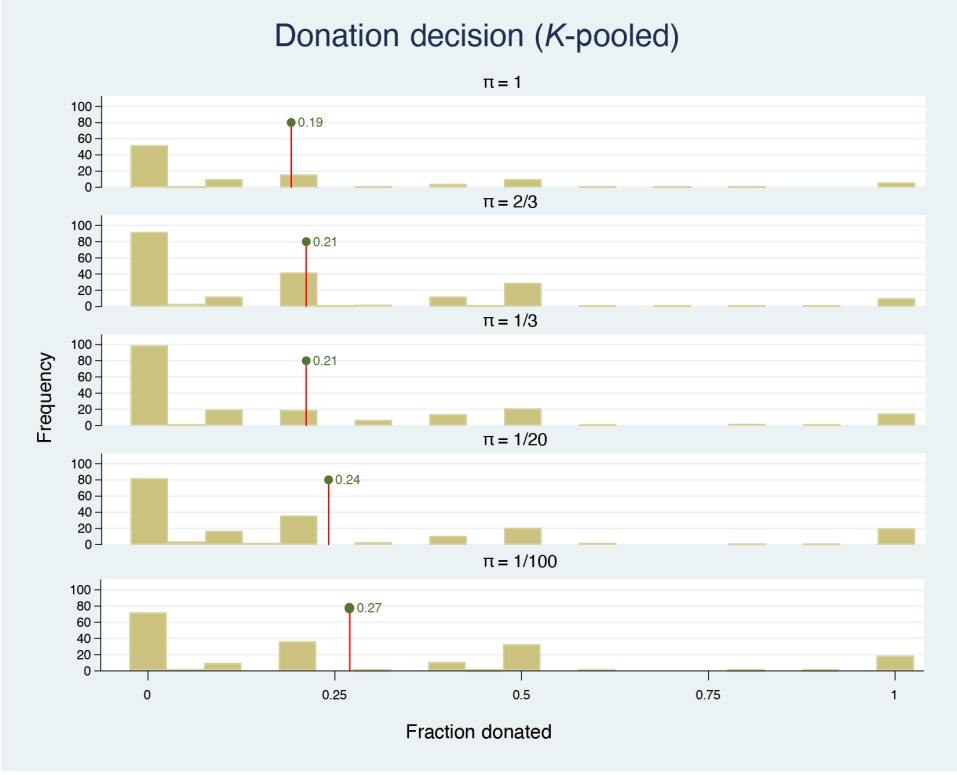

Fig. 9. Donation and π: Raw data (MTurk). The donation data from the MTurk experiment. The vertical lines indicate the mean donation of each treatment group.

Table 3. Donation and π: Linear regression (MTurk). This table presents regression results from the MTurk experiment. SEs in parentheses. Raw data shown in Fig. 9. Controls include indicator variables for gender, American, Indian, Christian, Atheist, aged 25 or younger, and aged 26 to 35, and continuous measures for religious attendance and accuracy in the lock-in data entry task. *$P < 0.10$, **$P < 0.05$, and ***$P < 0.01$.

	Ordinary least squares					
	(1)	(2)	(3)	(4)	(5)	(6)
	d^*		Expected income $E(\mathcal{X}_2)$		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)
κ 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

Table 4. Donation and π: Linear regression disaggregated by κ (MTurk). This table presents for four treatment groups the relationship between being consequential and the decision. Note: SEs in parentheses. *$P < 0.10$, **$P < 0.05$, and ***$P < 0.01$.

	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)		(0.0560)		(0.0630)		(0.0590)
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 to 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

Table 5. Donation and π: Nonparametric tests (MTurk). This table shows that, along most thresholds for π, Mann-Whitney tests yield significant differences in the distribution of donations as π increases.

	Wilcoxon-Mann-Whitney two-sided test (*P* values)		
	(1)	(2)	(3)
Thresholds	K-Unknown or 10¢	K = 0¢ or 50¢	K-Pooled
$\pi = 1$ versus $\pi \leq 0.67$	0.91	0.05	0.11
$\pi \geq 0.67$ versus $\pi \leq 0.33$	0.07	1.00	0.20
$\pi \geq 0.33$ versus $\pi \leq 0.05$	0.05	0.10	0.01
$\pi \geq 0.05$ versus $\pi = 0.01$	0.05	0.02	0.01
		π-Pooled	
$K \geq 10$¢ versus $K = 0$¢		0.040	
$K = 50$¢ versus $K \leq 10$¢		0.11	

d is then given by the following expression: If $\frac{1}{2} > d$, then $0 = \pi\lambda(2\beta - 1) + 2(\delta - d)$, else $0 = \pi\lambda(-2\alpha - 1) + 2(\delta - d)$.

The derivation is as follows: $\pi\lambda(1 - d - \alpha\max\{2d - 1,0\} - \beta\max\{1 - 2d,0\}) + (-(\delta - d)^2)$. This expression is quadratic in d, so the first-order condition, and hence moment conditions, will be linear in d. Thus, we estimate 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 $\pi\lambda(1 - d - \alpha(2d - 1)) + (-(\delta - d)^2)$. Thus, our linear regression is that, 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 the following general method of moments (GMM) condition

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

Thus, we run a linear regression of d on $1\left[\frac{1}{2} > d\right]\pi$ and $1\left[\frac{1}{2} \leq d\right]\pi$. We present estimates using two different instruments for $1\left[\frac{1}{2} \leq d_i\right]$, which results in similar point estimates (Table 8).

The bliss point is still 25%. Then, the first coefficient in the regression model indicates that while $d < 50\%$, donation increases as π decreases. However, once $d > 50\%$, donation decreases as π decreases. This switch is intuitive because the bliss point for duty is below 50% and we still assume the bliss point preferences by Cappelen *et al.* As π falls, they should move toward the bliss point, which is less than 50%. Our coefficients also have a structural interpretation for λ. Table 8 yields $\frac{\lambda(2\beta - 1)}{2} = -0.36$ and $\frac{\lambda(-2\alpha - 1)}{2} = 1.16$. Last, we need to make an assumption for α and β. For the range of plausible α and β values in Fehr and Schmidt (*3*), our data are inconsistent with the joint hypothesis of consequentialist motivations being Fehr-Schmidt, the duty motivation being bliss point,

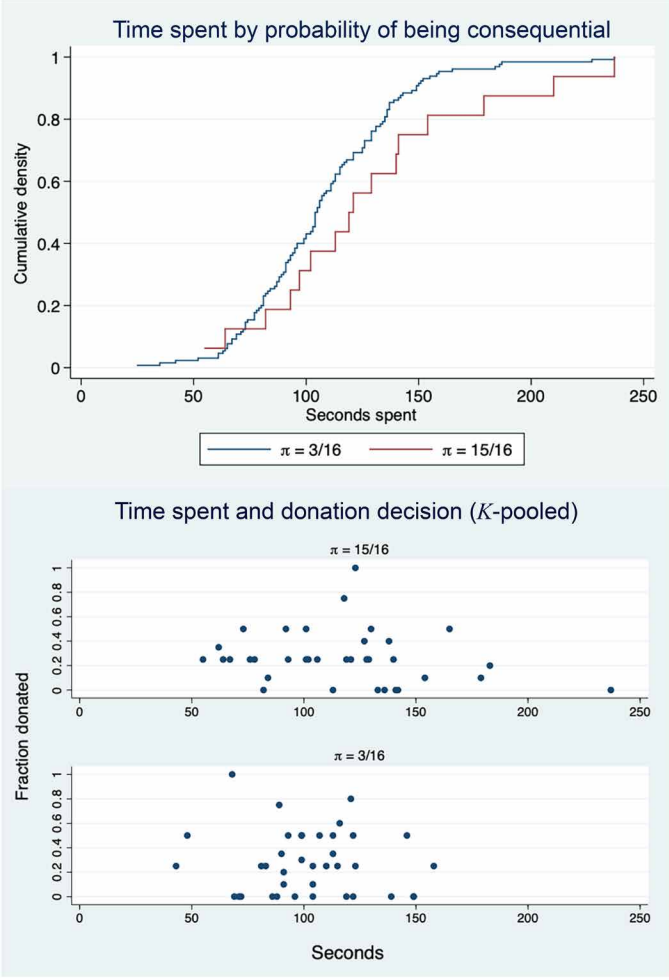

Fig. 10. Time spent (on donation decision): Laboratory. The cumulative density of time spent by probability of being consequential in the laboratory experiment. It also shows the relationship between donation and time spent.

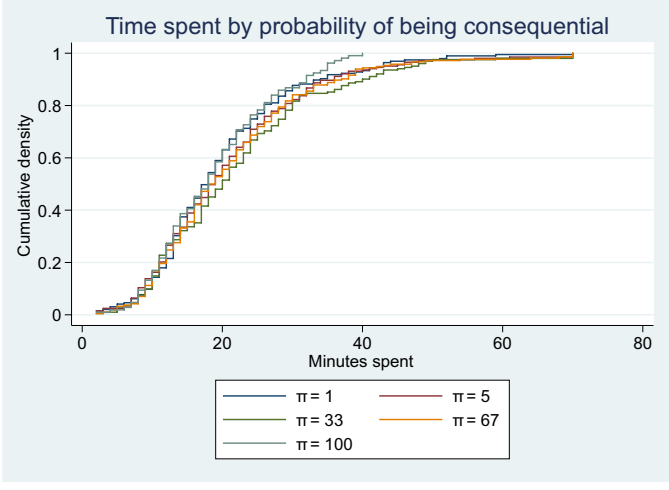

Fig. 11. Time spent (begin versus end time): MTurk. The cumulative density of time spent by probability of being consequential on the MTurk experiment.

Table 6. Time spent (begin versus end time): MTurk heterogeneity by $\frac{\delta d}{\delta \pi}$. Notes: SEs 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 SD of the first regression to account for uncertainty in the calculation of mixed-consequentialist score. Columns 3 and 5 use median regressions. *$P < 0.10$, **$P < 0.05$, and ***$P < 0.01$.

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)
π2		−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	

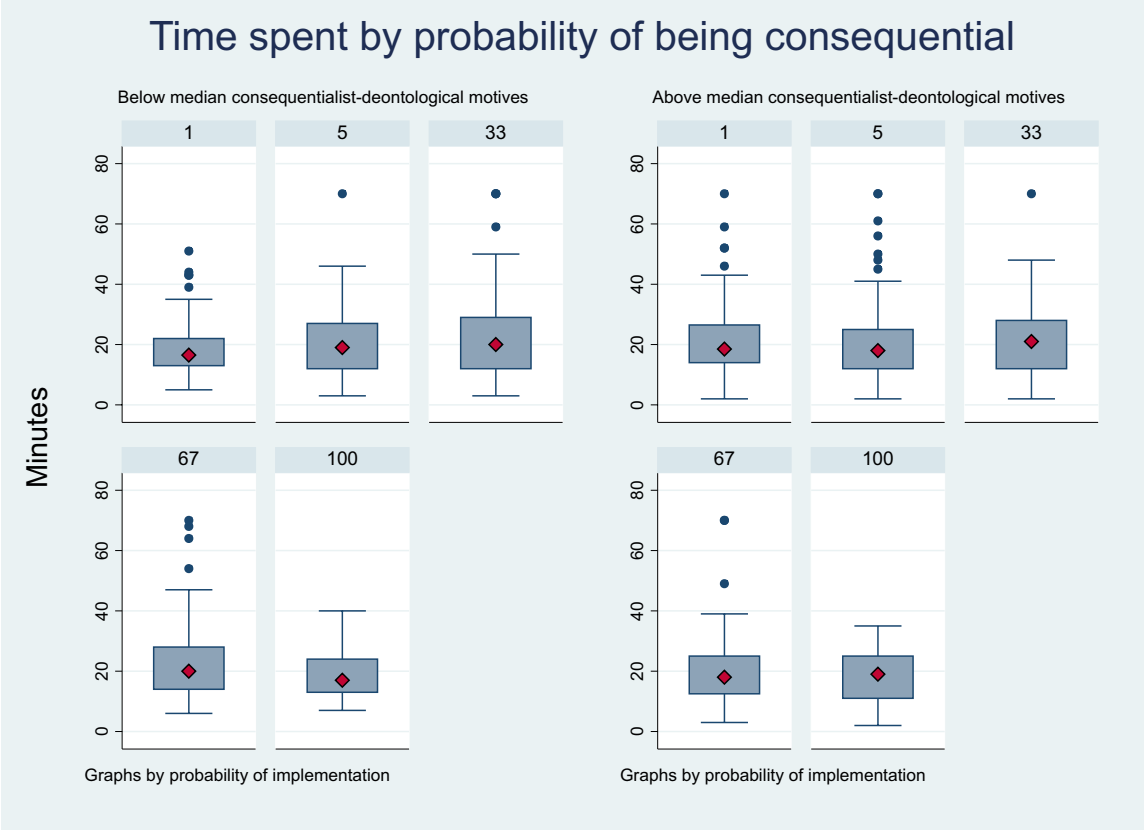

Fig. 12. Time spent by $\frac{\delta d}{\delta \pi}$: MTurk. The time spent as it varies by probability of being consequential for those who are categorized as mixed deontological-consequentialist. Red diamond, median.

Table 7. Who responds to π? (AMT). This table resents heterogeneity analysis of who is more responsive to the probability of being consequential. Notes: SEs 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 SD of the first regression to account for uncertainty in the calculation of mixed-consequentialist score. Columns 3 and 5 use median regressions. *$P < 0.10$, **$P < 0.05$, and ***$P < 0.01$.

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)
π2		−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	

Table 8. Donation and π: Linear regression. This table illustrates the structural identification strategy. Notes: SEs in parentheses. *$P < 0.10$, **$P < 0.05$, and ***$P < 0.01$. OLS, ordinary least squares; IV, instrumental variables.

	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

and a nonzero weight on consequentialist motivations. Together, each of the three exercises offer unique advantages and limitations that portray a picture of variance in response to the probability of implementation.

DISCUSSION

Recent advances in economic theory, motivated by experimental findings, have led to the adoption of models where individuals make decisions not solely based on self-interest (considering consequences for oneself) but also based on the consequences for others. Investigations of motives over decisions per se, independently of their consequences, are rare. Here, we formalize the notion of consequentialist and deontological motivations as properties of preference relations; we suggest and implement a thought experiment that uses revealed preference to detect deontological motivations—varying the probability that one's decision is consequential (i.e., implemented). For a consequentialist who satisfies FOSD, the optimal decision is independent of the probability that the action will be enacted. For a deontologist, the optimal decision is also independent of the probability. Only mixtures of both consequentialist and deontological motivations predict changes in behavior as the probability changes.

Our research design has some implications for the random lottery method in experimental economics. Prior formal observations support its use—roughly speaking, if individuals satisfy the independence axiom (*60*), then the random lottery method is valid—and these theoretical observations have been empirically validated (*61*, *62*). What we show is that when it comes to decisions that are not purely economic (e.g., social preference decisions that can have a deontological motive), if individuals satisfy FOSD, then the random lottery method can reveal different decisions that are more prosocial than when the decisions are consequential.

Future research may explore several legal applications. First, measuring 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) or did the shooter unintentionally commit the act of killing. In other instances, the law also cares about mental states beyond just the consequences, such as the litigant's motivations 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. More broadly, in equity law, judges may care about opportunistic behavior as opposed to the behavior itself, which is similar to the DM having both mens rea and actus reus. Last, some philosophers argue that human dignity derives from the possibility of deontological decision-making—"what commands respect is the capacity for morality" (*63*) and "Everything has either a price or a dignity. What has a price can be replaced by something else as its equivalent; what, on the other hand, is raised above all price and therefore admits of no equivalent has a dignity … humanity insofar as it is capable of morality is that which alone has dignity" (*1*).

MATERIALS AND METHODS

We ran the laboratory experiment in Zurich using zTree (*64*). We asked subjects aged 18 to 30 to make a donation decision out of an endowment of 20 Swiss francs (CHF) with the knowledge that we would shred their decision when it was not implemented. One session collected data from a classroom, but the procedures were the same and the endowment was 10 CHF. All our results are reported in terms of percent donation. The donation recipient was Doctors Without Borders as we believed this organization to be more salient in German-speaking countries.

Participants first saw a demonstration of a public randomization device (section S2 includes pictures and instructional materials) and a paper shredder; the shredding bin was opened to publicly verify that materials were truly going to be destroyed. Before the experiment, subjects were asked three IQ (intelligence quotient) tasks. If at least one answer was correct, then they proceeded to the donation decision and received information about their probability of implementation. We had a 2 × 2 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 nonconsequential state. The randomization wheel had 16 numbers. We only mentioned one or three of these numbers to the subject depending on their π. The 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.

After the wheel 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. We oversampled subjects who received low probabilities. If we assign the same number of subjects to each treatment condition, then far fewer data will be collected for $\pi = \frac{3}{16}$ treatment condition where only few envelopes are opened. We sought a roughly 1:1 ratio for the opened envelopes in the high and low π conditions. All results only analyze the decisions of envelopes opened as we do not have data for envelopes that were shredded.

We ran the online experiment using MTurk. We first asked MTurk subjects to transcribe three paragraphs of text to reduce the likelihood of their dropping from the study after seeing treatment. After the lock-in task, subjects have an opportunity to split a 50-cent bonus (separate from the payment they received for data entry) with the charitable recipient, the Red Cross. We believed the Red Cross to be more well known for MTurk subjects, who come mostly from the United States and India. Workers then provided their gender, age, country of residence, religion, and how often they attend religious services. We had 902 decisions from 902 subjects (two individuals did not report a complete set of demographic characteristics, so they are dropped in some of the regressions).

Participants were randomly assigned to one of five groups with π being 100, 66, 33, 5, and 1%. They were told in advance about the implementation probability. We randomized such that we collected roughly 200 subjects in each of the 66, 33, 5, and 1% treatments and 100 subjects in the 100% treatment. In addition, we randomize κ to be 50 cents (maximum) and 0 cents (minimum). Section S3 presents instructions. To assess potential anchoring effects induced by κ, we also ran an auxiliary experiment that randomized 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. We found that 18% of subjects gave 10 cents in the "κ = 10 cents" treatment, while 14% gave 10 cents in the "κ= unknown" treatment. Because we did not see significant anchoring effects, it is not the focus of our analysis. All our analyses are reported in terms of fraction donated from 0 to 1.

To estimate high and low $\frac{\delta d}{\delta \pi}$ and to explore sensitivity of the decision d to π, we construct synthetic cohorts. Formally, we estimate

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

We interpret the change in d to π as measuring the mixed consequentialist-deontological motives. We then compute for each individual

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

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

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abb3925

REFERENCES AND NOTES

1. I. Kant, Über ein vermeintes Recht aus Menschenliebe zu lügen. *Berlinische Blätter* **1**, 301–314 (1797).
2. M. Rabin, Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **83**, 1281–1302 (1993).
3. E. Fehr, K. M. Schmidt, A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817–868 (1999).
4. K. A. McCabe, M. L. Rigdon, V. L. Smith, Positive reciprocity and intentions in trust games. *J. Econ. Behav. Organ.* **52**, 267–275 (2003).
5. A. Falk, U. Fischbacher, A theory of reciprocity. *Games. Econ. Behav.* **54**, 293–315 (2006).
6. J. Dana, D. M. Cain, R. M. Dawes, What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organ. Behav. Hum. Decis. Process.* **100**, 193–201 (2006).
7. J. Dana, R. A. Weber, J. X. Kuang, Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Econ. Theory* **33**, 67–80 (2007).
8. R. Bénabou, J. Tirole, Incentives and prosocial behavior. *Am. Econ. Rev.* **96**, 1652–1678 (2006).
9. J. Andreoni, B. D. Bernheim, Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica* **77**, 1607–1636 (2009).
10. P. Foot, The problem of abortion and the doctrine of double effect. *Oxf. Rev.* **5**, 5–15 (1967).
11. F. J. Anscombe, R. J. Aumann, A definition of subjective probability. *Ann. Math. Stat.* **35**, 199–205 (1963).

12. J. Cilliers, O. Dube, B. Siddiqi, The white-man effect: How foreigner presence affects behavior in experiments. *J. Econ. Behav. Organ.* **118**, 397–414 (2015).
13. T. C. Bergstrom, R. J. Garratt, D. Sheehan-Connor, One chance in a million: Altruism and the bone marrow registry. *Am. Econ. Rev.* **99**, 1309–1334 (2009).
14. H. Choi, M. Van Riper, S. Thoyre, Decision making following a prenatal diagnosis of down syndrome: An integrative review. *J. Midwifery Womens Health* **57**, 156–164 (2012).
15. S. T. Trautmann, A tractable model of process fairness under risk. *J. Econ. Psychol.* **30**, 803–813 (2009).
16. M. W. Krawczyk, A model of procedural and distributive fairness. *Theor. Decis.* **70**, 111–128 (2011).
17. N. Chlaß, W. Güth, T. Miettinen, Purely procedural preferences-beyond procedural equity and reciprocity, *Tech. rep.*, (Stockholm School of Economics, Stockholm Institute of Transition Economics, 2014).
18. A. W. Cappelen, A. D. Hole, E. Ø. Sørensen, B. Tungodden, The pluralism of fairness ideals: An experimental approach. *Am. Econ. Rev.* **97**, 818–827 (2007).
19. A. W. Cappelen, J. Konow, E. Ø. Sørensen, B. Tungodden, Just luck: An experimental study of risk-taking and fairness. *Am. Econ. Rev.* **103**, 1398–1413 (2013).
20. J. Sobel, Interdependent preferences and reciprocity. *J. Econ. Lit.* **43**, 392–436 (2005).
21. G. A. Akerlof, R. E. Kranton, Economics and identity. *Q. J. Econ.* **115**, 715–753 (2000).
22. T. R. Tyler, The psychology of legitimacy: A relational perspective on voluntary deference to authorities. *Pers. Soc. Psychol. Rev.* **1**, 323–345 (1997).
23. A. Smith, *The Theory of Moral Sentiments* (A. Millar, 1761).
24. R. Bénabou, J. Tirole, Identity, morals, and taboos: Beliefs as assets. *Q. J. Econ.* **126**, 805–855 (2011).
25. W. H. Riker, P. C. Ordeshook, A theory of the calculus of voting. *Am. Polit. Sci. Rev.* **62**, 25–42 (1968).
26. T. Feddersen, S. Gailmard, A. Sandroni, Moral bias in large elections: Theory and experimental evidence. *Am. Polit. Sci. Rev.* **103**, 175–192 (2009).
27. M. Shayo, A. Harel, Non-consequentialist voting. *J. Econ. Behav. Organ.* **81**, 299–313 (2012).
28. S. DellaVigna, J. A. List, U. Malmendier, G. Rao, Voting to tell others. *Rev. Econ. Stud.* **84(1)**, 143–181 (2017).
29. I. Alger, J. W. Weibull, Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica* **81**, 2269–2302 (2013).
30. J. Andreoni, Impure altruism and donations to public goods: A theory of warm-glow giving. *Econ. J.* **100**, 464–477 (1990).
31. T. Ellingsen, M. Johannesson, Pride and prejudice: The human side of incentive theory. *Am. Econ. Rev.* **98**, 990–1008 (2008).
32. P. Battigalli, M. Dufwenberg, Guilt in games. *Am. Econ. Rev.* **97**, 170–176 (2007).
33. C. D. Batson, J. G. Batson, J. K. Slingsby, K. L. Harrell, H. M. Peekna, R. M. Todd, Empathic joy and the empathy-altruism hypothesis. *J. Pers. Soc. Psychol.* **61**, 413–426 (1991).
34. K. D. Smith, J. P. Keating, E. Stotland, Altruism reconsidered: The effect of denying feedback on a victim's status to empathic witnesses. *J. Pers. Soc. Psychol.* **57**, 641–650 (1989).
35. Z. Grossman, Self-signaling and social-signaling in giving. *J. Econ. Behav. Organ.* **117**, 26–39 (2015).
36. U. Gneezy, Deception: The role of consequences. *Am. Econ. Rev.* **95**, 384–394 (2005).
37. P. E. Tetlock, Thinking the unthinkable: Sacred values and taboo cognitions. *Trends Cogn. Sci.* **7**, 320–324 (2003).
38. S. Bowler, S. Polania-Reyes, Economic incentives and social preferences: Substitutes or complements? *J. Econ. Lit.* **50**, 368–425 (2012).
39. A. E. Roth, Repugnance as a constraint on markets. *J. Econ. Perspect.* **21**, 37–58 (2007).
40. N. G. Mankiw, M. Weinzierl, The optimal taxation of height: A case study of utilitarian income redistribution. *Am. Econ. J. Econ. Pol.* **2**, 155–176 (2010).
41. A. Falk, N. Szech, Morals and markets. *Science* **340**, 707–711 (2013).
42. T. Besley, Political selection. *J. Econ. Perspect.* **19**, 43–60 (2005).
43. L. Kaplow, S. Shavell, *Fairness Versus Welfare* (Harvard Univ. Press, 2006).
44. W. Sinnott-Armstrong, Consequentialism, in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. (The Metaphysics Research Lab, 2012).
45. J. Bentham, *Panopticon* (T. Payne, 1791).
46. K. J. Arrow, *Social Choice and Individual Values* (Cowles Foundation Monographs Series, Yale Univ. press, New Haven, ed. 3, 2012), Monograph 12.
47. L. Alexander, M. Moore, *Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. (The Metaphysics Research Lab, 2012).
48. C. Starmer, Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *J. Econ. Lit.* **38**, 332–382 (2000).
49. M. Friedman, L. J. Savage, The utility analysis of choices involving risk. English. *J. Polit. Econ.* **56**, 279–304 (1948).
50. L. J. Savage, *The Foundations of Statistics* (Courier Corporation, 1972).
51. R. Nozick, *Anarchy, State, and Utopia*, Harper Torchbooks (Basic Books, 1974).
52. M. J. Machina, "Expected utility" analysis without the independence axiom. *Econometrica* **50**, 277–323 (1982).
53. A. Tversky, D. Kahneman, Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).
54. J. Quiggin, A theory of anticipated utility. *J. Econ. Behav. Organ.* **3**, 323–343 (1982).
55. D. M. Kreps, *Notes on the Theory of Choice* (Westview Press Boulder, 1988).
56. D. Levhari, J. Paroush, B. Peleg, Efficiency analysis for multivariate distributions. *Rev. Econ. Stud.* **42**, 87–91 (1975).
57. N. T. Wilcox, Lottery choice: Incentives, complexity and decision time. *Econ. J.* **103**, 1397–1417 (1993).
58. D. G. Rand, J. D. Greene, M. A. Nowak, Spontaneous giving and calculated greed. *Nature* **489**, 427–430 (2012).
59. J. Bigenho, S.-K. Martinez, Social comparisons in peer effects, *Tech. rep.*, (UCSD, 2019).
60. C. A. Holt, Preference reversals and the independence axiom. *Am. Econ. Rev.* **76**, 508–515 (1986).
61. C. Starmer, R. Sugden, Does the random-lottery incentive system elicit true preferences? an experimental investigation. *Am. Econ. Rev.* **81**, 971–978 (1991).
62. J. D. Hey, J. Lee, Do subjects separate (or are they sophisticated)? *Exp. Econ.* **8**, 233–265 (2005).
63. J. Waldrom, How law protects dignity. *Camb. Law J.* **71**, 200–222 (2012).
64. U. Fischbacher, z-tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* **10**, 171–178 (2007).
65. J. Quiggin, Stochastic dominance in regret theory. *Rev. Econ. Stud.* **57**, 503–511 (1990).
66. P. Wakker, Savage's axioms usually imply violation of strict stochastic dominance. *Rev. Econ. Stud.* **60**, 487–493 (1993).
67. S. Choi, S. Kariv, W. Müller, D. Silverman, Who is (more) rational? *Am. Econ. Rev.* **104**, 1518–1550 (2014).
68. M. J. Machina, Dynamic consistency and non-expected utility models of choice under uncertainty. *J. Econ. Lit.* **27**, 1622–1668 (1989).

Acknowledgments: We thank research assistants and numerous colleagues at several universities and conferences. **Funding:** This project was conducted while D.L.C. received funding from the Alfred P. Sloan Foundation (grant no. 2018-11245), 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 Templeton Foundation (grant no. 22420). D.L.C. acknowledges IAST funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program (grant ANR-17- EUR-0010). This research has benefited from financial support of the research foundation TSE- Partnership and ANITI funding. **Ethics statement:** IRB approval was not required for this study by ETH Zürich. **Author contributions:** Both authors contributed equally to all parts of the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials..

Submitted 20 February 2020
Accepted 29 March 2022
Published 13 May 2022
10.1126/sciadv.abb3925

ScienceAdvances

Social preferences or sacred values? Theory and evidence of deontological motivations

Daniel L. ChenMartin Schonger

Sci. Adv., 8 (19), eabb3925. • DOI: 10.1126/sciadv.abb3925

Science Advances (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.