Ellingsen, Tore; Mohlin, Erik

Working Paper

A model of social duties

Working Paper, No. 2022:14

Provided in Cooperation with:
Department of Economics, School of Economics and Management, Lund University

Suggested Citation: Ellingsen, Tore; Mohlin, Erik (2022) : A model of social duties, Working Paper, No. 2022:14, Lund University, School of Economics and Management, Department of Economics, Lund

This Version is available at:
https://hdl.handle.net/10419/273644

Working Paper 2022:14

Department of Economics
School of Economics and Management

A Model of Social Duties

Tore Ellingsen
Erik Mohlin

August 2022

LUND
UNIVERSITY

A Model of Social Duties*

Tore Ellingsen† Erik Mohlin‡

Wednesday 24th August, 2022

Abstract

We develop a formal model of social duties. Duties to respect entitlements (duties of justice) differ from duties to promote well-being (duties of charity). A situation-specific version of our model takes entitlements as primitives. A fully portable version derives entitlements from situational characteristics. Utility functions obtain kinks where duties of justice and charity are exactly satisfied. Actions at these kinks are candidates for descriptive social norms. Empirically, duties are identified using Krupka-Weber appropriateness ratings, with negative ratings indicating entitlement violations. The model's predictions are confronted with established regularities and new survey evidence in seven pre-registered applications.

JEL Codes: D91, Z13

Keywords: Social norms, Social morality, Entitlements, Dutifulness, Charity

*This paper differs extensively from, yet replaces, earlier manuscripts entitled "Situations and Norms" and "Decency". Thanks to Matilde Casamonti for immaculate research assistance, and to Johannes Abeler, Sandro Ambuehl, Björn Bartling, Douglas Bernheim, Pol Campos-Mercade, Andrew Caplin, Vincent Crawford, Armin Falk, Ernst Fehr, Sebastian Fehrler, Gustaf Karreskog, Erik Gaard Kristiansen, David Laibson, Ulrike Malmendier, Karine Nyborg, Martin Oehmke, Robert Östling, Jon de Quidt, Zoltán Rácz, Alexandros Rigos, Felix Schafmeister, Christian Schultz, Peter Norman Sørensen, Mark Voorneveld, Christian Zehnder, and especially Klaus Schmidt and Roberto Weber, for helpful comments and discussions. Ellingsen gratefully acknowledges financial support from the Torsten and Ragnar Söderberg Foundation and from the the Norwegian Research Council (grant 250506). Mohlin gratefully acknowledges financial support from the Swedish Research Council (grant 2015-01751), Handelsbankens forskningsstiftelser (P19-0204), and the Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellowship 2016-0156).

†Affiliation: Stockholm School of Economics. Address: Department of Economics, Stockholm School of Economics, Box 6501, S—11383 Stockholm, Sweden. Email: gte@hhs.se

‡Affiliations: Lund University and the Institute for Futures Studies. Address: Department of Economics, Lund University, Tycho Brahes väg 1, 220 07 Lund, Sweden. E-mail: erik.mohlin@nek.lu.se.

1 Introduction

Why do people give to charity? Why do they tip? Who do they vote? Why do they pay taxes that they might easily have avoided? Why do they engage in social distancing in order to protect vulnerable strangers from virus infection? Why do they tell the truth when they could lie with impunity? In short, why are people so selflessly civil?[1]

One reason is *sympathy*. People are genuinely kind, taking pleasure from others' joy and pain from their suffering. Another reason is *duty*. People feel that they ought to act justly and charitably even if they incur material losses from doing so. There is already a rich array of models of sympathy and antipathy. For example, there are models of altruism and spitefulness (Edgeworth, 1881; Becker, 1974; Levine, 1998), fair-mindedness (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), and even a model that encompasses all of these motives (Charness and Rabin, 2002). Our objective is to build and evaluate a model of the duty motive that is of comparable precision and scope.

By modeling duty-based morality, we address the criticism of social preference theory that it is neglectful of context (Levitt and List, 2007), as sympathy is entirely a property of the individual. By contrast, individuals' social duties are determined at the group level.[2] In some societies, people have many social duties; in others, individual freedom is greater (e.g., Gelfand et al, 2011). In some societies, duties are mostly confined to the family, clan, or nation; in others, they are more universal (e.g., Enke, Rodríguez-Padilla, and Zimmermann 2020). Thus, the study of social duties has the potential to shed light on the large observed cross-culture differences in moral behavior (e.g., Cohn et al, 2019).

The central theme of our analysis is that duties come in two flavors. On the one hand are the duties to act rightly, what Cicero (44 BC) calls *duties of justice* (lat. *iustitia*) On the other hand are the duties to act well, what Cicero calls *duties of charity* (lat. *beneficentia*).[3] In the model, utility functions possess kinks at the ("reference") points where duties of justice and charity are exactly fulfilled. Such kinks naturally entail behavioral conformity at a particular unselfish action, a phenomenon that is otherwise challenging to explain (Bernheim, 1994). However, conformity requires that duties are commonly understood.

[1] We do not deny that some apparent civility is selfishness in disguise. People are sometimes afraid that a selfish act might hurt them by causing social contagion (Kandori, 1991) or by ruining their reputation (Sugaya and Wolitzky, 2021). They may even hold "magic beliefs" that if they fail to cooperate bad consequences will follow (Shafir and Tversky, 1992). Our assertion is merely that there *exists* civility in a sense that is truly separate from material self-interest.

[2] We shall not here discuss how duties are created and internalized. For complementary approaches to this question, see Casson (1991) and Akerlof and Kranton (2000). We also largely neglect the problem of competing loyalties.

[3] The modern literature on child development, taking its inspiration from Piaget (1932), studies behavior and emotions related both to justice (e.g., Kohlberg, 1964) and charity, or care (e.g., Gilligan, 1982). Recent work by social psychologists that is especially relevant to our analysis is that of Janoff-Bulman, Sheikh, and Hepp (2009) and Janoff-Bulman and Carnes (2013), who refer to justice as *proscriptive morality* and to care as *prescriptive morality*. For a broader survey of the literature on social morality, see Haidt (2008).

If people want to obey their duties, but have different understandings concerning what their duties are, the existence of duties is instead a source of non-conformity. We argue that such heterogeneous moral understandings may explain why neutrally framed laboratory experiments often generate a wide range of behavior along with contrasting moral defenses.

The distinction between justice and charity is tied to the concept of *entitlement*. Duties of justice entail respect for others' entitlements; duties of charity often (but not always) go beyond that minimal requirement. Sometimes, entitlements are explicit and obvious. For example, actions that violate entitlements in the form of well-defined property rights are understood by all to be proscribed. But what about the situations in which entitlements are not explicit? Often, the researcher can observe behavior, yet cannot directly observe entitlements. We therefore also build a portable version of our model, where all entitlements are derived exclusively from the game form. Concretely, we assume that payoff entitlements strike a balance between "might" and "right": People are entitled to a weighted average of morally ideal payoffs and payoffs resulting from purely selfish behavior.

Among other things, the portable formulation allows us to capture a classical context effect: the utility of a particular outcome is affected by what other outcomes that are available. More precisely, the model rationalizes the impact of unused options on people's moral choices that has been documented by List (2007) and Bardsley (2008).[4]

We measure perceptions of duties through the elicitation procedure devised by Krupka and Weber (2013) (KW). The KW-procedure elicits the social appropriateness of actions by inducing people to truthfully report what they consider to be the most common assessment of the social appropriateness of the various available actions. We define an action to violate a social proscription—i.e., to encroach on another's entitlement—if the action is classified as (somewhat or very) inappropriate. We complement the incentivized elicitation of social appropriateness with a non-incentivized elicitation of personal appropriateness (cf. Bašić and Verrina, 2021).[5]

To gauge the model's explanatory power, we confront it with a variety of evidence that sympathy-based models fail to accommodate. To corroborate the model's explana-

[4]Sen (1983,1993) prominently discusses the possibility that non-used options matter for people's moral choices. Still, decision theorists often continue to take consequentialism for granted (Hammond, 1996, footnote 3). Recent non-consequentialist theories include Dillenberger and Sadowski (2012), Saito (2015), and Evren and Minardi (2017), who all consider the case in which decision makers worry about what others will be thinking about them. Such concern for social esteem was previously modeled in less axiomatic fashion by, among others, Bénabou and Tirole (2006), Ellingsen and Johannesson (2008), and Andreoni and Bernheim (2009). By contrast, and like us, Cox et al (2019) consider entirely internalized morality and develop a model of observable reference points.

[5]Already Jasso and Opp (1997) construct similar ratings of personal appropriateness. However, they confine attention to binary decisions (participating in a political protest or not). While they note that an action can be either proscribed or prescribed, the binary setting does not admit that there can be a justice norm associated with an action that rates neutrally on their scale.

tion for the evidence, we also make several new KW-elicitations, one for each application. The seven phenomena that we address are:

1. *Giving.* Standard Dictator experiments typically yield gifts of zero and of half the available money, but also several other fractions between these two extremes (Engel, 2011). The model rationalizes both the extreme and the intermediate gifts, and so do the KW-elicitations.

2. *Giving and Roles.* In Dictator experiments, giving depends on how the roles were allocated and on other features of the past (e.g., Konow, 2000; Cherry, Frykblom, and Shogren, 2002; Kameda et al, 2002; Cappelen et al, 2007). The KW-elicitations support the hypothesis that it is considered justified to give less if the role of decider is allocated according to merit.

3. *Giving and Taking.* In Dictator experiments with taking options, the opportunity to take reduces the propensity to give even among people who do not avail themselves of the taking opportunity (List, 2007; Bardsley, 2008; Cappelen et al 2013b; Korenok, Millner, and Razzolini, 2014). KW-elicitations support the prediction that less generous giving is justified when taking is possible.

4. *Giving and Exit.* In Dictator experiments with unexpected exit options, the opportunity to exit is more attractive to subjects that gave more initially (Dana, Cain, and Dawes, 2006; Broberg, Ellingsen, and Johannesson, 2007; Lazear, Malmendier and Weber, 2012; DellaVigna, Malmendier, and List, 2012; Andreoni, Rao, Trachtman, 2017).

5. *Willful Ignorance.* In Dictator experiments with uncertain externalities, many subjects prefer to be ignorant rather than to learn the externality prior to acting (Dana, Weber, and Kuang, 2007; Feiler, 2014; Grossmann, 2014; Grossman and van der Weele, 2017; Freddi, 2021; and Serra and Szech, 2021).

6. *Incentive Paradox.* Fines sometimes backfire; they encourage the punished behavior instead of discouraging it (Gneezy and Rustichini, 2000). The theory implies that fines offer restitution and therefore make transgressions less immoral. KW-elicitations offer some support for this hypothesis.

7. *Lost Wallet Paradox.* People have a greater tendency to return found wallets to their owner when there is more money in them (Cohn et al, 2019). The theory rationalizes this behavior. KW-elicitations support the (weaker) hypothesis that it is morally more objectionable to keep wallets when they contain money than when they do not.

Items 1 and 3-5 were also central to the analysis of KW, which we build on. KW demonstrate that there is a reduced-form link between measures of appropriateness and behavior in these applications. Our complementary contribution is to develop and test a structural theory of the appropriateness of different behaviors. The final two items show that the model has applications far beyond conventional generosity experiments.

The moral philosophy of social duties has deep and durable roots. According to Marcus Tullius Cicero, whose book *On Duty* was written in year 44 BC, the social duties of justice and charity constitute the glue that holds societies together. Cicero's ideas have had a lasting influence on western moral thought. Major works of David Hume (1751) and Adam Smith (1759) devote considerable effort to explaining how the social duties of justice and charity are essential to societies' prosperity.[6] For example, Adam Smith argues in some detail that duties of justice and charity constitute more important drivers of prosocial behavior than does sympathy (Part III, Chapter V), and that justice is more important than charity (Part II, Section II, Chapter III). However, in the 20th century, economists largely ceased to study morality as a deep explanation for behavior,[7] leaving this topic to be explored by the other social sciences.[8] Indeed, as Granovetter (1985) documents, 20th century economists usually take one of two extreme positions, either assuming that moral concerns do not matter at all or that they impose binding constraints on behavior.[9] Our model takes a middle way, nesting the two extremes as special cases.[10]

The kinks of our utility function correspond to different notions of norms. The kink where responsibilities are exactly fulfilled corresponds to a *charity norm* of doing what is best (prescribed), whereas the kink where proscriptions are marginally avoided corresponds to a *justice norm* of not doing anything that is outright wrong (proscribed). Depending on the situation as well as on the distribution of dutifulness, either of these may be the best candidate for a *descriptive norm*, i.e., of a standard of behavior that many adhere to. In the formal literature on social norms, our model is most closely related to Ny-

[6] Instead of charity, Hume and Smith use the terms benevolence and beneficence respectively.

[7] When Camerer and Thaler (1995) argue against sympathy (altruism) as an explanation for unselfish behavior in the laboratory, it is perhaps telling that they choose the word "manners" rather than the more morally loaded "duties". That said, economists continued to use morally loaded terms when discussing purely *normative* criteria for good behavior; see Konow (2003) for a survey of formal models of moral ideals.

[8] Above all, the duty motive is integral to large parts of sociology, ever since Durkheim (1900) and Weber (1905). Legal scholars emphasize that laws have an *expressive function* and thus generate internalized norms (Sunstein, 1996; Cooter, 1998; Kahan, 2007). In political science, the duty motive remains widely accepted as a motivational force, explaining among other things why people vote (e.g., Riker and Ordeshook, 1968; Blais and Galais, 2016). For a recent perspective on social dutifulness from the perspective of evolutionary psychology, see Tomasello (2020).

[9] Granovetter's criticism is not confined to economics. He blames sociology for consistently taking an over-socialized approach to human behavior. Duesenberry (1960, p.233) expressed a similar sentiment with his famous quip: "Economics is about individuals' choices, sociology about how individuals don't have any choices to make."

[10] Within social psychology, our general approach has much in common with interdependence theory (Kelley and Thibaut, 1978; Rusbult and Van Lange, 2008), but as it is both more formal and more specific, it is more amenable to testing.

borg (2000), Brekke, Kverndokk, and Nyborg (2003), Bicchieri (2005), López-Pérez (2008), and Huck et al (2012), although none of these make an analogous distinction between justice norms and charity norms.[11]

2 A Simple Model of Social Duties

A single decision-maker makes a decision that impacts herself and one other person. We refer to the decision-maker as Decider and the other person as Other. For the most part, we take for granted that Decider has duties toward Other. That is, we do not model the determination of moral boundaries or how the duty arose.[12]

Decider has access to a set of actions $\mathcal{A}$, with typical element a. Other has no action to take. Actions entail material consequences for both persons. Let $\mathcal{M} \subset \mathbb{R}^2$ be the set of feasible consequences, with $\mathcal{M}_i$ denoting the set of feasible consequences for Person i, and let $x : \mathcal{A} \to \mathcal{M}$ denote the outcome function. That is, the pair $(x_d(a), x_o(a))$ represents the material consequences to Decider (d) and Other (o) from action a.

As a minimal running example, suppose Decider can choose to Help or Not help, represented as $\mathcal{A} = \{H, N\}$, with material consequences $x(H) = (0, 3)$, and $x(N) = (1, 1)$. This Helping situation is illustrated in Figure 1.

H	$0, 3$
N	$1, 1$

Figure 1. Helping situation

We next describe the two kinds of social duties. The first is the duty to respect others' interests (duties of justice). The second is the duty to promote the community's objectives (duties of charity).

2.1 Duties of justice

An action is either morally permitted or morally proscribed (forbidden). We say that Decider has a *duty of justice* not to take proscribed actions. Let $\mathcal{J}$ denote the set of just (non-proscribed) actions. Proscriptions are closely linked to Other's payoff-entitlement, henceforth simply called the *entitlement*; an action is proscribed if and only if the action entails a violation of Other's entitlement.

[11] Other seminal work on social norms has pursued different goals and therefore either (i) lacks powerful testable implications, like Stigler and Becker (1977) and Akerlof and Kranton (2000) (Sobel, 2005, demonstrates the close connection between them), (ii) is confined to particular applications, like Kandel and Lazear (1992) and Lindbeck, Nyberg, and Weibull (1999), or (iii) explores departures from full rationality, like Rabin (1994, 1995), and Konow (2000), and Gächter and Riedl (2005, 2006).

[12] Extending the model to allow endogenous duties, for example due to contracting or other forms of communication, requires extensive-form notation and is the topic of a separate paper.

In some situations entitlements are naturally viewed as being derived from proscriptions, which are taken as exogenous. In other situations proscriptions are derived from entitlements. In the latter case we may either take the entitlements as primitives, or derive them from the game form.

In situations covered by general rights, entitlements are derived from proscriptions.[13] Concretely, Other's entitlement is the payoff that she can secure herself when Decider takes a just action:

Definition 1 *When proscriptions are exogenous, Other's entitlement is*

$$e_o = \min_{a \in \mathcal{J}} x_o(a). \tag{1a}$$

For example, In the Helping situation, our running example, $e_o = 3$ if only H is just and $e_o = 1$ if both H and N are just. If Decider takes action N when only H is just, the infringement is 3-1=2.

Conversely, for given entitlements, we define just actions as follows.

Definition 2 *When proscriptions are endogenous, an action a is perceived as just by Decider if $x_o(a) \geq e_o$.*

That is, Decider considers an action proscribed if it gives Other less than Other's entitlement, and just otherwise.

In situations with endogenous proscriptions our model comes in two versions. A situation specific version takes entitlements as primitive and recover them from data (as descrived further below). A more portable version derives entitlements from the structure of the situation, striking a balance between a moral ideal and the power possessed by the parties.[14] Formally, let the Decider's *selfish* action be $a^{\text{ego}} = \arg\max_a x_d(a)$, inducing the outcome $(x_d^{\text{ego}}, x_o^{\text{ego}}) = x(a^{\text{ego}})$. (In all our applications there is a unique selfish action.) Let a^{ideal} be a morally ideal action, inducing the outcome $(x_d^{\text{ideal}}, x_o^{\text{ideal}}) = x(a^{\text{ideal}})$. We define the moral ideal below.

With this notation, we can define entitlements as a weighted average of the ideal allocation and the Decider's selfish allocation,

$$e_o^* = \beta x_o^{\text{ideal}} + (1 - \beta) x_o^{\text{ego}}, \tag{1b}$$

[13] For example, in most societies, there are property rights related to prior possession, as documented by Curry, Mullins, and Whitehouse (2019). For discussions of prior possession principles in the "lawless" context of the Californian gold rush, see Umbeck (1977) and Zerbe and Anderson (2001). In the context of social duties, Cicero (44 BC, Book 1, Paragraph 21) and especially Hume (1751, Section III, Part II) devote considerable attention to the origin and wisdom of private property rights. Sugden (1986) presents a related evolutionary-game theoretic account.

[14] For discussions of how such an entitlement might arise evolutionarily, see, e.g., Binmore (2005). For a somewhat related discussion of justice norms in sociology, see, e.g., Stolte (1987).

where $\beta \in [0,1]$ is the weight on "right" (the ideal outcome) relative to "might" (the selfish outcome).

However, e_o^* is not always feasible. To illustrate, suppose H is the morally ideal action in the Helping situation. Then $e_o^* = 3\beta + (1-\beta) = 1 + 2\beta$. This payoff is only attainable for Other if β is either 0 or 1. If e_o^* is not feasible, Other's perceived entitlement is instead given by the feasible consequence closest to the point e_o^*. If two feasible consequences are equally close, the tie is broken in favor of Decider. That is, when unique, the entitlement is

$$e_o = \arg \min_{x_o \in \mathcal{M}_o} |x_o - e_o^*|. \tag{1c}$$

When the solution is is not unique, pick the solution with the highest x_d. In the Helping situation, we thus have $e_o = 3$ if $\beta > 1/2$ and $e_o = 1$ if $\beta \leq 1/2$.

Regardless of whether entitlements are taken as primitive or derived, the *harm* to Other (as viewed by Decider) is defined as the payoff deficit relative to the entitlement,

$$h(a) = \max\{0, e_o - x_o(a)\}. \tag{2}$$

In the Helping situation, if Other is entitled to get 3 and Decider plays N, the harm is $3 - 1 = 2$. If instead Other is entitled to get 1 and Decider plays H, the harm is $\max\{0, 1-3\} = 0$.

2.2 Duties of charity

Duties of charity are defined in relation to the community's objectives. Suppose the community's objective is to pursue efficiency and equality, so that communal value might be written [15]

$$c(x) = x_d + x_o - \alpha \left| x_d - x_o \right|, \tag{3}$$

where $\alpha \geq 0$ is the weight on equality.[16] Let $\hat{a}$ be a maximizer of $c(x(a))$, let $\hat{\mathcal{A}}$ denote the set of such *communally ideal* actions, and let

$$s(a) = c(x(\hat{a})) - c(x(a)) \tag{4}$$

be called the *shortage* produced by action a.

In the Helping situation, the communally ideal action depends on α. If $\alpha \leq 1/3$, then $s(H) = 0$ and $s(N) = 1 - 3\alpha$. If $\alpha > 1/3$, then $s(H) = 3\alpha - 1$ and $s(N) = 0$.

From now on, we assume that the morally ideal action corresponds to a communally

[15] In the case of $n > 2$ players, the more general expression is $c(x) = \sum_i x_i - \alpha \sum_j \left| x_d - \sum_j x_o/n \right|$. Naturally, there are other principles of distributive justice that could also come into play.

[16] A generalization would be to let α depend on relative payoffs. In this way, one could incorporate additional insights from models of inequity aversion.

ideal action.[17]

2.3 Blameworthiness and behavior

The blameworthiness of an action is given by its contribution to harm and shortage. We take the view that duties of justice are always important, whereas duties of charity are more context-dependent (c.f. Cicero, Book I, Paragraphs 40-59). Accordingly, we write blameworthiness as[18]

$$b(a) = h(a) + \gamma s(a). \tag{5}$$

We assume that there are situations in which charity is salient and situations in which it is not—either because care conflicts with justice or for other reasons. When charity is not salient, we set $\gamma = 0$.

We assume that Decider's utility function is

$$u_d(a) = x_d(a) - \delta b(a). \tag{6}$$

We refer to the individual-level parameter $\delta \geq 0$ as Decider's *dutifulness* and to $\delta b(a)$ as Decider's *guilt*.[19] Note that people comply with social duties *only* because they have internalized them. We do not consider compliance that is caused by future rewards or punishments. Nor do we consider that people perform duties in order to gain social esteem, as in Bernheim (1994), or to protect others from being disappointed, as in Charness and Dufwenberg (2006). In this respect, our model is a close cousin of DellaVigna, List, and Malmendier (2012), whose additive and linear formulation we also adopt.[20]

When $\delta = 0$, Decider is *homo oeconomicus*, a selfish materialist whose behavior is unaffected by obligations and responsibilities. As δ grows large, Decider becomes *homo sociologicus*, whose choices are essentially determined by the society. Like Granovetter (1985), we are interested in the intermediate case.

[17] This is a good assumption whenever the ideal action is just—which is going to be the case in all our applications here. But one can easily think of cases where the communally ideal action violates Other's entitlement. For example, suppose you are lost in the mountains and find a cabin. What are the circumstances under which it's morally defensible to break in? Or what are the circumstances under which Robin Hood's behavior is defensible?

[18] A natural alternative formulation is $b(a) = (1-\gamma)h(a) + \gamma s(a)$. However, that yields more complicated expressions.

[19] We interpret Kimbrough and Vostroknutov (2016) as providing evidence that dutifulness is a personality trait with predictive power across situations. Note that any difference in the average value of δ across societies can be interpreted as differences in the weight that societies attach to duty.

[20] A less tractable but possibly more realistic alternative is to let guilt be convex in blame and to add a fixed cost for all violations obligations. Abeler et al (2019) propose this functional form for the disutility of lying. Comparing our blameworthiness function b to the "social pressure cost" formulation of DellaVigna, List, and Malmendier (2012), the main difference is that we impose more structure through h and s; the kink at $h(a) = 0$ is particularly important.

3 Measuring Proscriptions and Prescriptions

The distinction between proscriptions and prescriptions is a key feature of the model. Obtaining a credible empirical measure of this distinction is therefore essential for building confidence in the model's mechanism.

3.1 Elicitation of Social and Personal Appropriateness

Krupka and Weber (2008, 2013) define a measure that is very well suited for our purposes. Their procedure elicits the social appropriateness of actions by inducing people—either participants or spectators—to truthfully report what they consider to be the group's most common assessment of the social appropriateness of the various available actions. The KW-scale runs from "very socially inappropriate" to "very socially appropriate". We empirically identify an action as violating an obligation if the action is classified as inappropriate. Such an operationalization of violations of proscriptions is consistent with Krupka and Weber's (2008, Appendix) instruction to participants. Our nearly identical formulation runs as follows:

> By socially appropriate, we mean behavior that most people agree is the "correct" or "proper" or "ethical" thing to do. Another way to think about what we mean is that if a person were to select a socially inappropriate choice, then someone else might be angry at the person for doing so.

Inspired by Bašić and Verrina (2021), we also elicit personal (as opposed to social) appropriateness ratings for a separate set of subjects. These subjects were simply asked to report their own personal view of the appropriateness of the different actions, without any monetary incentives. Except for this difference the phrasing of the instructions were kept as similar as possible to the instructions for the social appropriateness elicitation.[21]

> By appropriate, we mean the behavior that you personally would consider to be the "correct" or "proper" or "ethical" thing to do. Another way to think about what we mean is that if a person were to select an inappropriate choice, then you might be angry at the person for doing so. We are interested in your personal opinion, independently of the opinion of others.

3.2 Experimental procedures

Our experiments were conducted in the online labor market Prolific during October and November 2021.[22] In total we used 2000 subjects, with 150-200 subjects in each treatment.

[21] To the best of our knowledge we are the first to compare KW elicitation with an un-incentivized question about personal appropriateness; Bašić and Verrina (2021) do not incentivize the social appropriateness questions.

[22] For a comparison of various online labor markets, see Peer et al (2021).

Before each experiment, we pre-registered our hypotheses with Open Science Foundation. There are three pre-analysis plans; see the Online Appendix Section 4 for links.

Subjects were all based in the United States, and were informed about this fact. Depending on the expected duration of the experiment, each subject received a participation reward of either 2 or 3 GBP, with the hourly pay falling in the interval 22-27 GBP (amounting to 29-36 USD at the time). For details, see Online Appendix 4.

Each subject is asked to rate the appropriateness of behavior in several situations. In order to maintain subjects' attention throughout the session while controlling for spillovers across situations each subject is randomly allocated to a subset of situations, the order of which is also randomized.

Subjects are either asked to guess the most common social appropriateness rating or to provide their own personal rating of appropriateness. In the social appropriateness treatments, participants earn either 2.5 or 4 GBP extra if they match the most common appropriateness rating when one of their ratings is randomly drawn for payment. In the personal appropriateness treatments, participants only earn the participation reward.

Our instructions (see Online Appendix) closely follow the format of Krupka and Weber (2008, Appendix). A key difference is that we include the neutral option "neither appropriate nor inappropriate" to their categories "very inappropriate", "somewhat inappropriate", "somewhat appropriate", and "very appropriate".

Like Krupka and Weber, we also assign numerical value -1 to the lowest category and $+1$ to the highest, and place the other categories equidistantly in this interval, implying that the neutral option has value 0. Throughout we report the results of pre-registered t-tests comparing average appropriateness ratings across survey questions.[23]

4 Applications: Addressing Seven Puzzles

We next confront the model with some well-known puzzles. In each case, we first briefly describe the puzzle. We then translate the situation that generates the puzzle into the language of our model and derive the model's predictions. Finally, we display the relevant new data that we have collected and discuss to what extent the data support the model's interpretation of the original puzzle.

Our first five applications are variations of minimally framed Dictator experiment. They overlap substantially with the applications considered by Krupka and Weber (2013), though we look at them from the perspective of our structural theory.

The last two applications concern two rather different field experiments, which both

[23] We have made the same comparisons with Wilcoxon-Mann-Whitney tests (available upon request). The results are similar and never overturn our conclusions from the t-tests. There is no evidence of gender differences (results from pre-registred tests available upon request). There are some minor order effects in the dictator experiments with taking options, see Online Appendix.

challenge existing theories. One considers the impact of fines at a daycare center and the other considers the choice to return a lost wallet to the rightful owner.

We summarize the key predictions of our models in *observations*, sometimes under simplified parameter assumptions. In the Online Appendix we provide complete results and proofs.

4.1 Standard Dictator Experiments

In a standard Dictator experiment, the experimenter has given Decider an endowment which she can share however she likes with Other (usually called the recipient). For concreteness, let the endowment be 10 dollars. For mnemonic reasons—and comparability with the experimental literature—we call the action g (for giving) rather than using the generic label a. That is, Decider picks a gift g in the interval $[0, 10]$ while keeping $k = 10 - g$ for herself.[24]

According to Equation (1b), Other's entitlement (as perceived by Decider) is

$$e_o = \beta \cdot 5 + (1 - \beta) \cdot 0 = 5\beta. \tag{7}$$

Thus, Decider's utility is

$$\begin{aligned} u_d &= 10 - g - \delta\left(h(g) + \gamma s(g)\right) \\ &= 10 - g - \delta\left(\max\{0, e_o - g\} + \alpha\gamma|(10 - g) - g|\right) \\ &= 10 - g - \delta\left(\max\{0, 5\beta - g\} + \alpha\gamma|10 - 2g|\right), \end{aligned}$$

where the first equality uses (6) and (5), the second equality uses (2), (3), and (4), and the last equality uses (7). Maximizing this piece-wise linear objective function with respect to g subject to the constraint that $g \leq 0$ yields the following result.[25]

Observation 1 *In a standard Dictator experiment, Decider gives*

$$g = \begin{cases} 0 & \text{if } \delta < \dfrac{1}{1 + 2\alpha\gamma}; \\ 5\beta & \text{if } \delta \in \left(\dfrac{1}{1 + 2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right); \\ 5 & \text{if } \delta > \dfrac{1}{2\alpha\gamma}. \end{cases} \tag{8}$$

The intuition is straightforward and illustrated in Figure 2. If Decider is sufficiently dutiful (e.g., $\delta = 3$ in Figure 2), she will maximize communal value and share equally. If

[24] In experiments, the subjects may be confined to pick integer amounts, but here we stick with the classical formulation. The Appendix expresses the analysis in terms of k rather than g.

[25] Note that the cases are not exhaustive: if $\delta = 1/(1 + 2\alpha\gamma)$ then any $g \in [0, 5\beta]$ is optimal, and if $\delta = 1/(2\alpha\gamma)$ then any $g \in [5\beta, 5]$ is optimal.

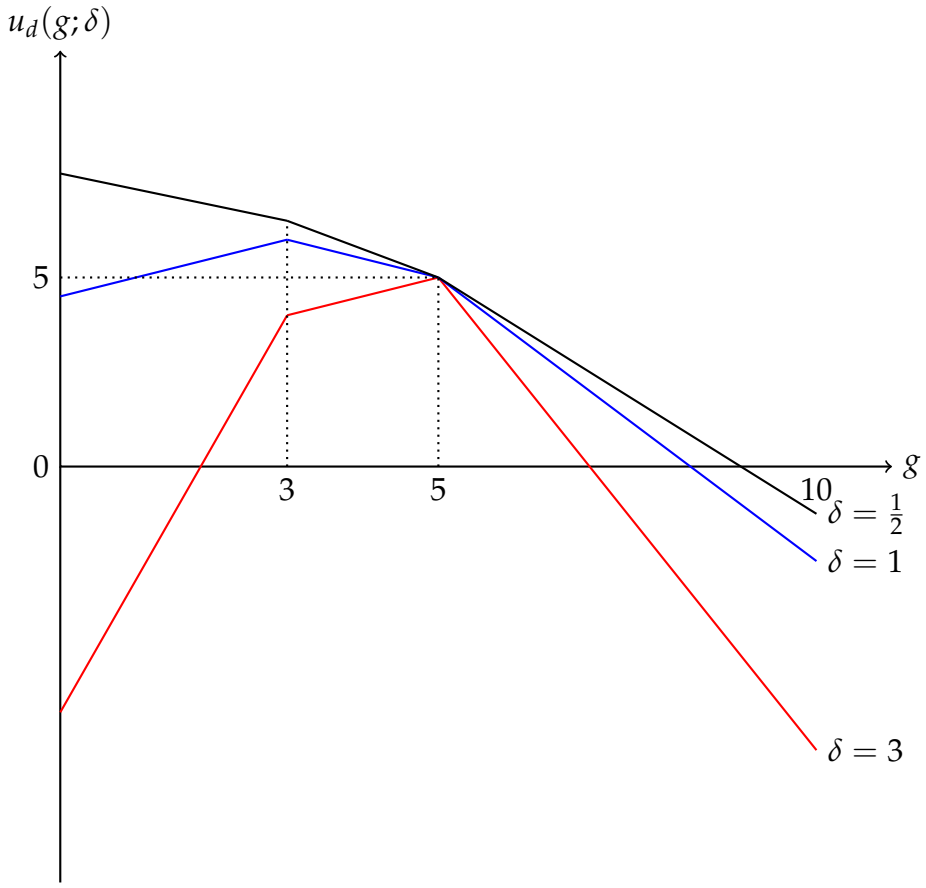


Figure 2. Decider's utility in the Dictator experiment ($\alpha = 1/4, \beta = 3/5, \gamma = 1$)

her dutifulness is sufficiently low (e.g., $\delta = 1/2$), she will neglect both duties and gives nothing. For an intermediate range of dutifulness (including $\delta = 1$), Decider neglects the prescription but obeys the proscription and gives Other exactly his entitlement. Depending on the value of β, Other's entitlement may be at 0, 5, or somewhere in between.[26]

We think that this is a plausible account of available data, which displays two modes at 0 and 5, and where giving 2 and 3 is also not uncommon. Our interpretation is that these gifts primarily correspond to the deciders' conception of the recipients' entitlement, i.e., to a non-degenerate distribution of β. Among those that give 5, the model says that there can be both those who put a weight close to 1 on the moral ideal (β close to 1) and those that are highly dutiful (have a large δ) and therefore maximize social appropriateness.

This interpretation is supported by the elicited appropriateness rankings of Krupka and Weber (2013, Table 1). To ensure comparability across our different treatments, we replicate their elicitation as part of our experiment; see Table 1.

[26]The parameters α and γ only enter through the product $\alpha\gamma$. This feature is due to the fact that all actions a entail the same efficiency. Our first four applications have this feature, whereas the last three do not.

Action	VSI	SI	N	SA	VSA
Give $ 0	0.62	0.20	0.08	0.07	0.03
Give $ 1	0.43	0.28	0.11	0.16	0.02
Give $ 2	0.27	0.40	0.10	0.20	0.02
Give $ 3	0.11	0.41	0.18	0.26	0.05
Give $ 4	0.03	0.22	0.23	0.43	0.10
Give $ 5	0.01	0.01	0.12	0.25	0.62
Give $ 6	0.04	0.07	0.13	0.42	0.35
Give $ 7	0.05	0.14	0.20	0.27	0.34
Give $ 8	0.08	0.20	0.16	0.18	0.38
Give $ 9	0.14	0.14	0.16	0.18	0.38
Give $ 10	0.20	0.10	0.16	0.14	0.41

Table 1. Dictator experiment: social appropriateness ratings

0	1	2	3	4	5
0.18	0.11	0.06	0.14	0.26	0.23

Table 2. Dictator experiment: recipient's entitlements

The even split is the only action that a majority considers to be very socially appropriate, with many of these also finding it very socially appropriate to give more than half. Let us now infer what the implied entitlements are. With the exception of a few subjects who think that an even split is inappropriate—here rounded up to 2 percent—all the subjects agree that the recipient is not entitled to more than $5. But widespread agreement ends there. About a quarter of the subjects (0.03+0.22) consider that it is very or somewhat inappropriate to give $4. In other words, a fraction 0.25-0.02=0.23 think that the recipient is entitled to exactly $5. Analogous computations yield the distribution of perceived entitlements in Table 2.[27]

It may seem surprising that people can have such different views about entitlements. Is it really plausible that a fifth of the subjects think that the recipient is entitled to half and almost as many think that the recipient is entitled to nothing? A separate source of evidence on this question comes from Ellingsen and Johannesson (2008), who let the recipient write a message to the decider after observing the Decider's decision. The messages' emotions range from profuse gratitude to great anger. While gratitude is more

[27]The computations are: A fraction (0.11+0.41-0.23-0.02)=0.26 think that the entitlement is $ 4, a fraction (0.27+0.40-0.26-0.23-0.02)=0.14 think that the entitlement is $3, a fraction (0.43+0.28-0.26-0.23-0.14-0.02)=0.06 think that the entitlement is $2, a fraction (0.62+0.20-0.26-0.23-0.14-0.06-0.02)=0.11 think that the entitlement is $1, and the remaining 0.18 think that the recipient is not entitled to anything. (We arrive at the latter number either as (1-0.23-0.26-0.14-0.06-0.11-0.02) = 0.18 or more directly as the fraction of subjects considering "Keep $ 10" either "neutral", "somewhat appropriate" or "very appropriate"—i.e., (0.08+0.07+0.03)=0.18.

common after large gifts and anger is more common after small gifts, the overlap is striking. Some recipients express gratitude for small gifts, apparently thinking that they are not entitled to them.[28]

Perhaps such disagreement is exactly what we should expect. The Dictator experiment is an unfamiliar setting. How strong is the Decider's property right? Should the Decider consider the endowment to be entirely common or should she consider it the same way as her other money? The unfamiliarity of the Dictator situation makes it difficult to generalize findings about donation levels to settings outside of the laboratory.

4.2 Dictator Experiments and Roles

Even if people hold heterogeneous views about the level of entitlement, Dictator experiments can still be useful for studying the general determinants of entitlements. In fact, the sensitivity of donations to contextual factors may be quite revealing. For example, Konow (2000) and Cappelen et al (2007) use Dictator experiments to demonstrate that people are more reluctant to share earned endowments than lucky endowments. The authors believe that this difference in behavior is caused by a difference in perceived entitlements, and hence the perceived duty to share. (By contrast, Cherry, Frykblom, and Shogren, 2002, offer a rather different explanation for this finding that is based on rationality and attention, so the entitlement hypothesis is not vacuous.) In order to test their hypothesis, we here elicit social and personal appropriateness ratings under the different endowment regimes. We compare (i) the standard Dictator experiment, where the Decider role is allocated by chance, with (ii) a setting in which the Decider role as allocated based on performance on a *quiz*, and (iii) a setting in which the Decider role is allocated by chance but has to *produce* the endowment that she can divide, by performing a real effort task. Figure 3 reports our findings. Observe in particular how it is considered much less inappropriate to keep almost everything when the endowment is "earned".[29] The average social appropriateness rating of keeping everything is -0.35 in the quiz setting and -0.21 in the production setting, compared with -0.65 in the standard Dictator experiment. For each of the options of giving 0, 1, or 2, the differences in (both social and personal) appropriateness rating between the standard experiment and the two alternatives are significant at $p < 0.0001$ (t-test, pre-registered hypotheses).[30]

[28] The messages are available at shorturl.at/uAW08. Here are excerpts from two contrasting messages from recipients who both got SEK 20 out of the endowment SEK 120. Recipient 1: "Thanks for the money! You made a generous choice by giving me twenty." Recipient 2: "You greedy bastard. Normally you give at least 1/3. You have to be a boy!"

[29] The downward movement in average ratings does not cause any reduction in heterogeneity. In fact, we find the opposite. People disagree even more regarding the appropriateness of low donations when the endowment is earned.

[30] A related issue concerns the sensitivity of behavior to framing manipulations; some authors have portrayed Dictator experiments as being potentially highly sensitive to minute changes in labeling of actions or the situation as a whole. Addressing this worry, Dreber et al (2013) find to the contrary that pure labeling

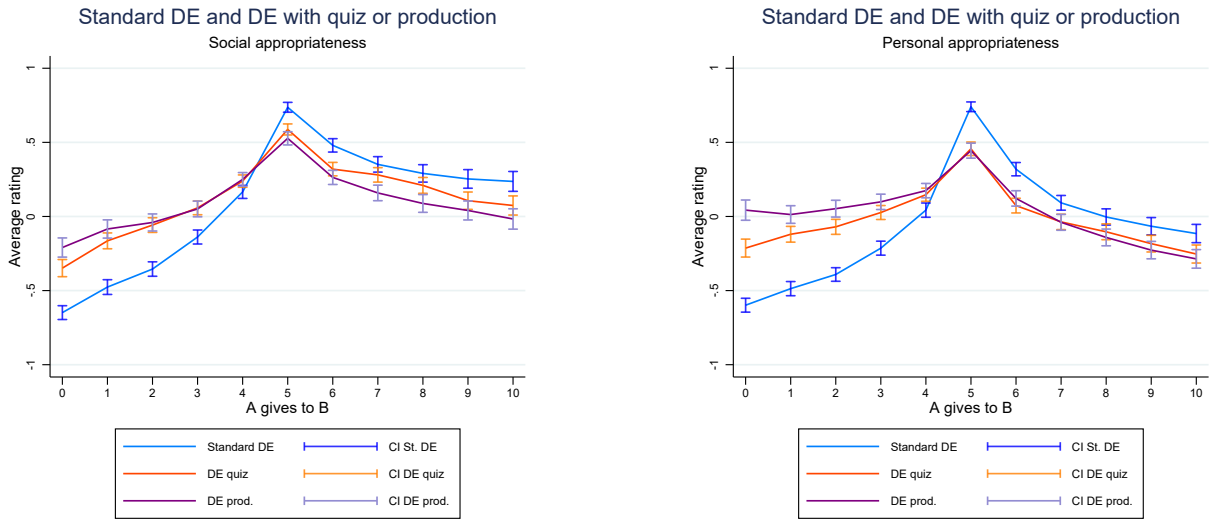


Figure 3. Average social (right) and personal (left) appropriateness ratings in the standard DE, DE with quiz, and DE with production (95% confidence intervals)

4.3 Giving and Taking Experiment

List (2007) and Bardsley (2008) conduct a Dictator experiment in which the availability of an option to take (steal) influences subjects' willingness to give. The most striking finding is that the taking option does not merely shift behavior away from keeping the full initial endowment towards taking—positive gifts also become less common and smaller. This is not predicted by any of the existing sympathy-based models of social preferences that define utility over material allocations. Another striking finding is that not all who give nothing in the standard treatment will take everything in the take-treatments.

For easy comparison, we transform the 5 dollar Dictator experiment considered by List to a 10 dollar experiment and consider two treatment variations in addition to the baseline. In all treatments both Decider and Other are allocated 10 dollars, and Decider is given an additional 10 dollars to divide between the two of them. In one treatment, Decider can take up to two dollars from Other, and in a second treatment she can take up to 10 dollars. Call them Take 2 and Take 10 respectively.

In our framework, taking option directly affects entitlements. Recall that $e_j = 5\beta$ in the baseline situation. Once it becomes possible for Decider to take 10 dollars, Other's new implicit entitlement is

$$e_o^T = \beta \cdot 5 + (1-\beta) \cdot (-10) = 15\beta - 10.$$

The framework also allows (but does not imply) that context matters in an additional

effects are minuscule in Dictator experiments (but see the "Bully" treatment of Krupka and Weber, 2013).

way. The taking option may highlight to the participants the general proscription against stealing. This focal proscription—corresponding to $\beta = 0$—might replace the implicit entitlement the participants have previously been inferring. Thus, participants who previously have experienced that small gifts are unjust might now only experience that they are uncharitable, and that $e_o = 0$.

Solving the utility-maximization problem as before (Equation (8)), but with the two alternative entitlements, yields the following solution.

Observation 2 *In the Dictator experiment with an additional option to take 10, Decider's choice is*

$$g = \begin{cases} -10 & \text{if } \delta < \dfrac{1}{1+2\alpha\gamma}; \\ 15\beta - 10 & \text{if } e_o = e_o^T \text{ and } \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right); \\ 0 & \text{if } e_o(i) = 0 \text{ and } \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right); \\ 5 & \text{if } \delta > \dfrac{1}{2\alpha\gamma}. \end{cases} \tag{9}$$

Comparing to (8), we see that Decider types that gave nothing in the standard treatment now take all they can. Those that gave according to the recipient's entitlement also give less than before, as the entitlement has gone down. However, we potentially have a new group that give exactly nothing, namely some of those who (now, but not before) consider that only taking is proscribed. In summary, for those that held $\beta = 1$ and do not perceive the recipient's entitlement to drop to zero with the taking option there is no revision, but all other types revise downwards. Given the previously inferred distribution of β, some of the latter will now be net takers.[31]

To interpret the findings, we again replicate KWs elicitations (Krupka and Weber, 2013, Figure 5). Figure 4 displays the data. It yields support for each of our two mechanisms. In support of the moving-implicit-entitlement hypothesis, the whole rating distribution moves to the left as taking options increase. In support of the focal-no-steal hypothesis, the ratings move steeply from inappropriate to appropriate around the gift of 0 when taking options are available. For each of the options of giving 0, 1, or 2, the differences in (both social and personal) appropriateness rating between the standard experiment and the two alternatives are significant at $p < 0.0001$ (t-test, pre-registered hypotheses).

[31] One feature of List's data that is not captured by our results is that there are more people who take as much as they can when they can take up to 10 than when they can take up to 2. This pattern can be explained by adding a fixed cost of positive harm, as we demonstrate in the Online Appendix.

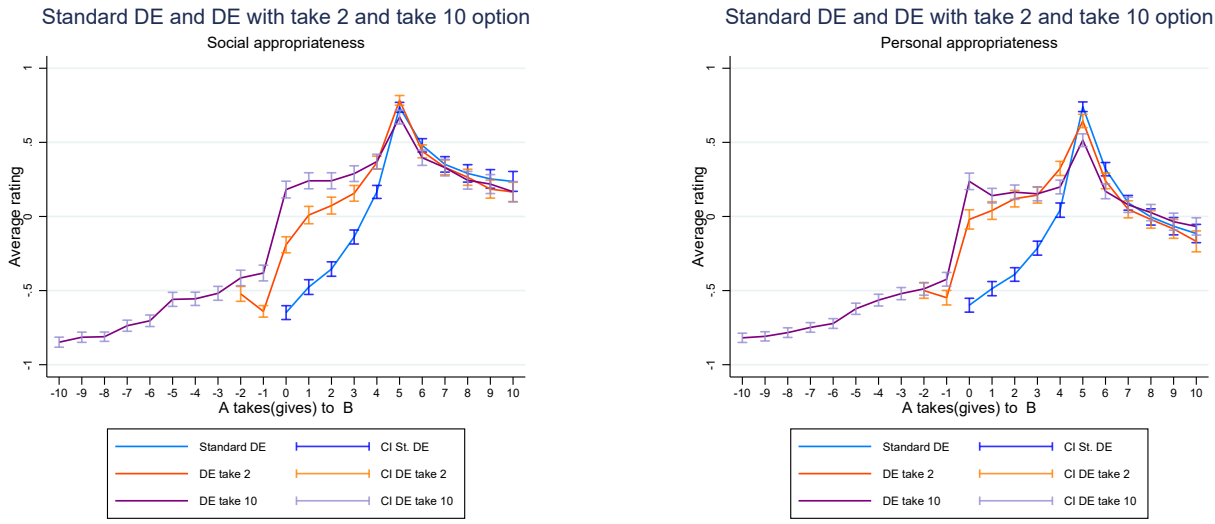


Figure 4. Average social (right) and personal (left) appropriateness ratings in the standard DE and DE with take options (95% confidence intervals)

4.4 Exit Experiment

In an experiment devised by Dana, Cain, and Dawes (2006), subjects are initially informed that they are in a Dictator situation. But after having made the allocation choice, Decider is told that Other is not yet aware of the experiment. Decider is given the option to exit for a price of 1. In the case of exit, Decider thus keeps 9, and Other will never be informed. About a third of the subjects choose to exit, a finding that was replicated and elaborated—with even greater exit rates—by Broberg, Ellingsen, and Johannesson (2007). There is also a tendency that generous sharers are more likely to exit, a finding that is further corroborated in closely related work by Lazear, Malmendier, and Weber (2012).

As noted by Dana, Cain, and Dawes (2006), exiting is inconsistent with sympathy-based social preference models in which utility is determined by material allocations, since exiting implements an allocation $(9,0)$ that is more unequal or less efficient than an allocation that could have been implemented without exiting, such as the allocation $(9,1)$ or the allocation $(10,0)$. However, according to our duty-based model, exiting is potentially easy to rationalize. For example, suppose that there is no duty to refrain from exiting.[32] In that case, subjects will compare the utility of getting 9 dollars to the utility given by the best choice according to Equation (8). Straightforward computations reveal that the outcome is as follows.

Observation 3 *In the Dictator experiment with an exit option, and there is neither an obligation*

[32] If there is a duty not to exit, the original argument applies.

nor a responsibility to remain, Decider's final choice (disregarding indifferences) is

$$g = \begin{cases} 0 & \text{if } \delta < \min\left\{\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{5(\beta+2\alpha\gamma)}\right\}; \\ 5\beta & \text{if } \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right) \text{ and } \delta < \dfrac{1-5\beta}{10\alpha\gamma(1-\beta)}; \\ Exit & \text{otherwise.} \end{cases} \tag{10}$$

Thus, comparing (10) to (8), we see that when there is no duty to remain, exiting with 9 dollars is the best option for all subjects that were originally giving one dollar or more.[33] Even some of those who were originally keeping everything are willing to sacrifice a dollar in order to exit. They can exit without feeling any guilt, but will feel guilty if they stay.

What if Decider perceives a duty to remain in the original situation? In the case of the Dictator experiment with taking options, we argued that the introduction of the taking option may signal to Decider that Other's entitlement is reduced to zero. Similarly, the presence of the exit option may nullify the perceived entitlement of Other. This means that exiting creates no harm, but still creates shortage.

Observation 4 *In the Dictator experiment with an exit option, if there is a responsibility but no obligation to remain, then Decider's final choice (disregarding indifferences) is*

$$g = \begin{cases} 0 & \text{if } \delta < \min\left\{\dfrac{1}{1+2\alpha\gamma}, \dfrac{1-5\beta}{5\beta-1+\alpha\gamma}\right\}; \\ 5\beta & \text{if } \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right) \text{ and } \delta < \dfrac{5\beta-1}{1-\alpha\gamma+10\alpha\gamma\beta}; \\ 5 & \text{if } \delta > \dfrac{1}{2\alpha\gamma}; \\ Exit & \text{otherwise.} \end{cases} \tag{11}$$

Now, all who initially set $g = 5$ remain. Those who initially set $g < 5$ remain if their dutifulness is sufficiently low. There are individuals who initially chose $g = 0$ and then prefer to exit if

$$\delta \in \left(\frac{1-5\beta}{5\beta-1+\alpha\gamma}, \frac{1}{1+2\alpha\gamma}\right).$$

A sufficient condition for this interval to exist is $\beta > 1/5$. Thus, when there is a responsibility to remain, it is no longer true that those that gave most initially are also most keen to exercise the exit option.

[33] Furthermore, if $1 < 8\alpha\gamma + 5\beta$ then

$$\frac{1-5\beta}{10\alpha\gamma(1-\beta)} < \frac{1}{1+2\alpha\gamma},$$

meaning that everyone who initially chooses $g > 0$ exits.

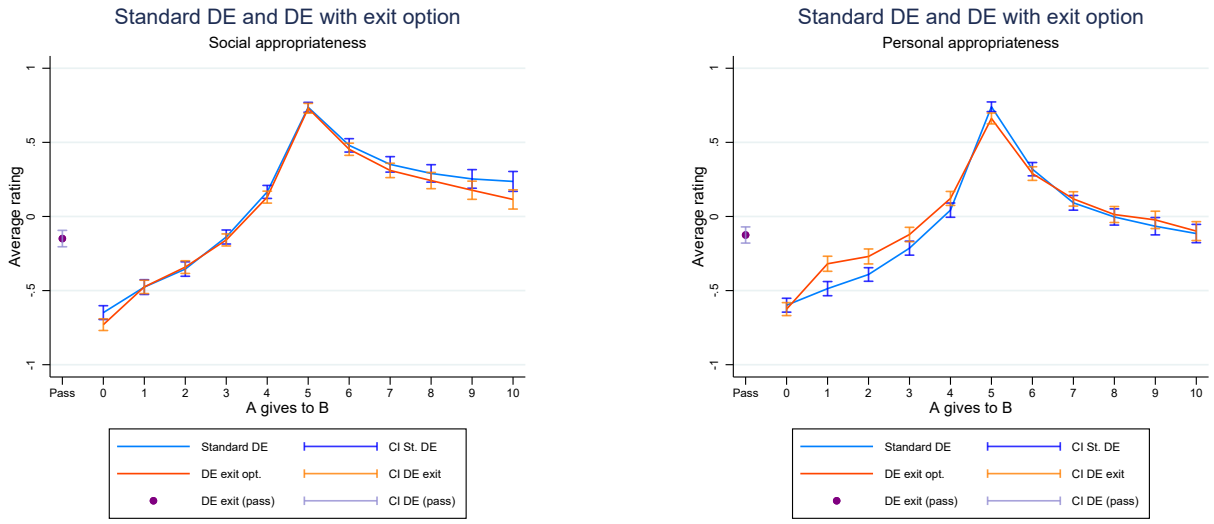


Figure 5. Average social (right) and personal (left) appropriateness ratings in the standard DE and DE with exit option (95% confidence intervals)

In our experiment, we elicit KW-ratings for the original experimental setting of Dana, Cain, and Dawes (2006), again following in the footsteps of Krupka and Weber (2013, Figure 3). Figure 5 reports our finding. The average social appropriateness rating of not exiting and keeping 9 is -0.63, whereas the average social appropriateness rating of exiting (and leaving with 9) is -0.13. The difference is significant at $p < 0.0001$ (t-test, pre-registered). The same us true for personal appropriateness.

As the curves demonstrate, compared to the standard Dictator experiment, the appropriateness of the various donations is virtually the same. Moreover, just as in Krupka and Weber (2013), the exit option is rated to be almost neutral on average, but with the average masking great individual variation, as shown in Table 3. Roughly half the participants consider exit to be either very or somewhat inappropriate, about 20 percent consider it neutral, and the remaining 30 percent consider exit to be either somewhat (17 percent) or very (13 percent) appropriate. These elicitations provide a clear reason why not everyone exits; half of the raters think that Deciders have an obligation to refrain from exiting, just as they have an obligation to give a non-negative amount. On the other hand, at least 13 percent of the raters consider that remaining is not a duty at all.[34]

Overall, the distribution of appropriateness ratings seems quite consistent with the observed heterogeneous behavioral pattern. There are enough people judging the exit option appropriate to justify a sizable exit rate, but also enough people considering exit inappropriate to justify that as many or more refrain from exiting.

[34]The number could be considerably higher, because it is only a duty to remain if remaining is considered more appropriate than whatever donation the person has made.

Action	VSI	SI	N	SA	VSA
Exit	0.24	0.26	0,20	0.17	0.13
Give $ 0	0.67	0.21	0.06	0.03	0.03
Give $ 1	0.37	0.38	0.11	0.11	0.03
Give $ 2	0.21	0.48	0.12	0.17	0.02
Give $ 3	0.08	0.42	0.26	0.21	0.03
Give $ 4	0.03	0.21	0.03	0.37	0.08
Give $ 5	0.01	0.01	0.12	0.23	0.63
Give $ 6	0.02	0.08	0.20	0.38	0.32
Give $ 7	0.05	0.14	0.23	0.29	0.29
Give $ 8	0.09	0,17	0.22	0.21	0.32
Give $ 9	0,17	0,16	0.17	0.16	0.34
Give $ 10	0,23	0,12	0.20	0.10	0.35

Table 3. Dictator experiment with exit: social appropriateness ratings

4.5 Information Avoidance Experiment

Dana, Weber, and Kuang (2007) (DWK) provide another important objection to sympathy-based consequentialist social preference models.[35] They demonstrate that many people prefer not to know whether there is a reason to be charitable or not, in an apparent violation of the independence axiom. In their experiment, Decider chooses between two actions, A and B. The actions generate a known material payoff to Decider, whereas the payoff to Other depends on the state. More precisely, the payoffs are:

- State 1 (non-aligned) payoffs: $A = (6, 1), B = (5, 5)$.
- State 2 (aligned) payoffs: $A = (6, 5), B = (5, 1)$.

In the Baseline treatment, the state is known to Decider, and most subjects choose action B in State 1 and A in State 2. The more interesting treatment is the Hidden Information treatment. There, Decider is not informed about the state, but told that both states are equally likely. Decider is given the choice between privately revealing the state before choosing the action, or to take the action without knowing the state. That is, Decider now takes two actions. The first action is whether to reveal or not. The second action is either A or B. Altogether, Decider has six strategies. Let us denote them $(RAA, RAB, RBA, RBB, NA, NB)$, where RAA denotes "reveal, then play A in both states" and NB denotes "not reveal, then play B" and so on.

In the Hidden Information treatment of DWK's experiment almost half of the subjects choose NA. That is, they remain ignorant and pick the selfish action A. The remainder

[35] See also, among others, Bartling, Engl, and Weber (2014), Feiler (2014), Grossmann (2014), Grossman and van der Weele (2017), Freddi (2021), and Serra and Szech (2021) for further evidence about strategic moral ignorance both in laboratory settings and in practice.

mostly play RBA. The behavior NA is incompatible with consequentialist preferences, according to which Decider would always want to play RBA. More precisely, under such preferences, NA violates the independence axiom.[36] By contrast, as we shall now demonstrate, the duty-based model does admit NA.

In order to analyze the non-revelation decisions, we must make an assumption regarding the blameworthiness of NA and NB.

actions that generate uncertain harm and shortage. We assume that blameworthiness is proportional to expected harm and shortage. For the revelation decisions, we assume that $\beta > 1/2$, so that Other's entitlement is 5 in both states.[37]

Let us illustrate the nature of the computations by considering two strategies, NA and RAA.

- NA: With probability $1/2$, A is a non-ideal action. In the non-aligned state, NA generates a shortage of efficiency amounting to $(5+5)-(6+1)=3$ and a shortage of equality amounting to $(6-1)-0=5$, so the total shortage is $s(NA)=3+5\alpha$. Since there is no obligation to reveal, $h(NA)=0$.

- RAA: Shortage is the same as under NA. Given revelation, there is an obligation to play B in the nonaligned state, so in this state $h(NA)=4$.

Table 4 summarizes all the material payoffs, Other's harm, the community's shortage, and Decider's utility for all six strategies. There are only two undominated choices, NA

	Non-aligned			Aligned			
k=	x_d, x_o	h	s	x_d, x_o	h	s	u_d
RAA	$6,1$	4	$3+5\alpha$	$6,5$	0	0	$6-\frac{1}{2}\delta(4+\gamma(3+5\alpha))$
RAB	$6,1$	4	$3+5\alpha$	$5,1$	4	$5+3\alpha$	$5.5-4\delta(1+\gamma(1+\alpha))$
RBA	$5,5$	0	0	$6,5$	0	0	5.5
RBB	$5,5$	0	0	$5,1$	4	$5+3\alpha$	$5-\frac{1}{2}\delta(4+\gamma(5+3\alpha))$
NA	$6,1$	0	$3+5\alpha$	$6,5$	0	0	$6-\frac{1}{2}\delta\gamma(3+5\alpha)$
NB	$5,5$	0	0	$5,1$	0	$5+3\alpha$	$5-\frac{1}{2}\delta\gamma(5+3\alpha)$

Table 4. Information avoidance experiment

and RBA. NA dominates RAA and RBA dominates the remaining strategies.

[36] Any theory of expected utility defined over final outcomes is unable to explain that someone who chooses $B=(5,5)$ over $A=(6,1)$ in the un-aligned state, and chooses $A=(6,5)$ over $B=(5,1)$ in the aligned state also chooses not to reveal the true state of the world and chooses A. To see this, note that if $u(6,5)>u(5,1)$ and $u(5,5)>u(6,1)$ then the independence axiom implies $pu(6,5)+(1-p)u(5,5)>pu(6,5)+(1-p)u(6,1)$.

[37] The case $\beta<1/2$ yields $e_o=1$. Then RAA and NA are equally bad, and we cannot explain a preference for NA. See the Online Appendix for additional analysis.

Observation 5 *Suppose $\beta > 1/2$. Then, Decider plays RBA if*

$$\delta > \frac{1}{\gamma(3+5\alpha)}$$

and NA if the inequality is reversed.

As claimed, the model's prediction is thus consistent with the behavioral evidence.[38]

The most interesting feature is that NA is not dominated by RBA. The key assumption for generating that result are that it is not an obligation to reveal and that β is large enough to make RBA is the most responsible choice. As before, a way to evaluate our assumptions is to elicit appropriateness ratings. While the published version of Krupka and Weber (2013) does not address the information avoidance experiment, the earlier draft Krupka and Weber (2008) does. The elicitation reveals that it is considered inappropriate to take the selfish action when the state is known to yield a low payoff for the opponent. However, it is not generally inappropriate to take the selfish action when the impact on the opponent's payoff is unknown; on average this action is considered neutral. In other words, there is not a general obligation to be informed, just as we assumed. Is there a social responsibility to be informed? Again the answer is affirmative. According to Krupka and Weber (2008), RBA obtains a much higher social appropriateness score than the neutral score obtained by NA.

We have also made our own elicitations. Figure 6 reports the averages.

Our data replicate all the key findings from Krupka and Weber (2008): On average, there is no obligation to reveal; non-revelation is somewhat appropriate (average social appropriateness of non-revelations is 0.11 when paired with action A and 0.16 when paired with action B). It is highly appropriate to reveal, find that the state is non-aligned, and play the unselfish action (average social appropriateness 0.82). By contrast, it is inappropriate to reveal, find that the state is non-aligned, and play the selfish action (average social appropriateness -0.52). In other words, the average ratings justify our above assumptions. The disaggregated data show that there is a minority of about 20 percent that considers it somewhat (13 percent) or very (4 percent) inappropriate not to reveal the state; remarkably, these fractions are almost identical irrespective of whether

[38] Feiler (2014) examines the effect of changing the probabilities of the different states in the Hidden Information experiment. She finds that the share of subjects that choose to become informed is decreasing in the probability p of the aligned state. A slight generalization of our model can account for this counterintuitive finding. To do so we need to assume that blame is convex in the sum of harm and shortage. As with the linear specification, the undominated strategies are RBA and NA with

$$u_d(RBA) = 5 + p > 6 - \delta(3(1-p))^2 = u_d(NA) \iff \delta > \delta^*(p) := \frac{1}{9(1-p)}.$$

Note that the threshold $\delta^*(p)$ is increasing in p implying that fewer reveal when p increases. The material payoff of RBA is increasing linearly in p, whereas the blame from expected harm and shortage associated with RBA is convex in p.

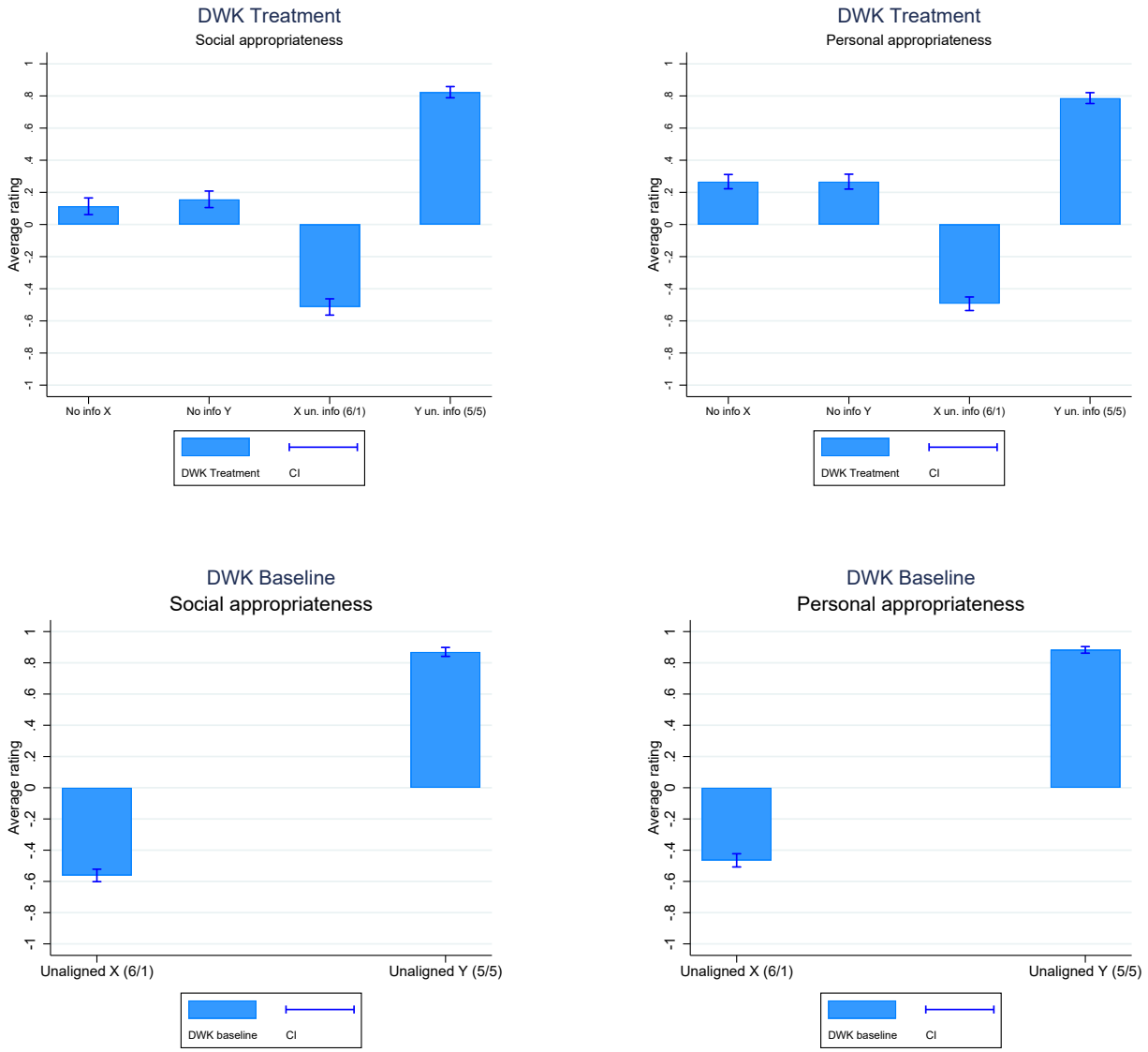


Figure 6. Information avoidance experiment: Average social (right) and personal (left) appropriateness ratings in Treatment (top) and Baseline (bottom) of DWK (95% confidence intervals)

a	x_d, x_o	h	u_d
S	$-k, 0$	0	$-k$
N	$0, -d$	d	$-\delta d$

Table 5. Daycare experiment, without fine

the the action choice following non-revelation is A (best for Decider) or B (best for Other).

We now turn to two rather different applications.

4.6 Incentive Paradox: The Daycare Experiment

A famous field-experiment by Gneezy and Rustichini (2000) documents that monetary incentives can backfire. The study considers the impact of imposing a penalty on parents who pick up their children late from daycare. If they are more than ten minutes late, parents need to pay a penalty corresponding to about eight US dollars in today's value. Instead of encouraging greater punctuality, the penalty prompted more parents to pick up late. As we shall now see, this outcome is consistent with the model under the mild assumption—which we corroborate shortly—that picking up late is considered proscribed.

Think of the initial situation between a parent and the daycare center workers. Let d denote the cost to the workers from a delayed parent and let k be the parent's material cost of being punctual. Let S denote "sacrifice" (being punctual) and let N denote "no sacrifice" (being late), and suppose the latter is considered proscribed, and the former is not.

To simplify, set $\gamma = 0$. That is, this situation is governed only by the daycare center's right to punctuality, not by concerns about efficiency or distribution. The situation is summarized by Table 5. The penultimate column shows that the action N is associated with a harm of d. The final column displays the parent's utility. Let $\delta(0)$ denote the threshold level of dutifulness, above which $u_d(S) > u_d(N)$. From Table 5, we see that

$$\delta(0) = k/d.$$

Suppose now that the daycare center introduces a fine $f < d$ for being late. Furthermore, for now, suppose that the perception of the day care center workers' entitlement is unaffected by this. For simplicity we assume that the workers receive this fine. Still, since f does not fully compensate the workers' loss, being late is still considered proscribed. Thus, the new situation is captured by Table 6. Note that the fine has two opposing effects on the parent's utility. On the one hand, it represents a loss of money, but on the

a	x_d, x_o	h	u_d
S	$-k, 0$	0	$-k$
N	$-f, f-d$	$d-f$	$-f-\delta(d-f)$

Table 6. Daycare experiment, with fine

other hand it represents a reduction in guilt. The new dutifulness threshold is

$$\delta(f) = (k-f)/(d-f).$$

Observe that the threshold coincides with $\delta(0)$ when $f = 0$. We say that the fine f backfires if $\delta(f) > \delta(0)$, i.e., if parents must be more dutiful in order to be punctual. Observation 6 states the result of this comparison.

Observation 6 *Suppose that the introduction of a fine does not affect entitlements. A penalty $f < d$ can make a parent less punctual if and only if $k > d$.*

In other words, a parent whose cost of being punctual exceeds the workers' benefit from punctuality will become less punctual when there is a fine. The intuition runs as follows. A parent who is punctual despite $k > d$ must have a high degree of dutifulness, and hence would feel considerable guilt when imposing harm on the workers by not being punctual. Since the fine serves to compensate the workers, it alleviates the dutiful parent's guilt to such an extent that the guilt reduction outweighs the material cost of paying the fine. The fine offers an avenue for *restitution*.

In order for there to be an aggregate decrease in punctuality, parents with $k > d$ must thus constitute a large enough fraction of those parents who are late to begin with.

Above we assumed that the introduction of the fine did not affect the perception of the day care center workers' entitlement. Alternatively, it may be that the introduction of the fine changes the way the parents construe the situation, and induces them to believe that being late is not proscribed—justifying action N. This is similar to the notion that taking options in the Dictator experiment can make the no-stealing norm salient, nullifying different entitlements. Indeed, this is one of several theories that Gneezy and Rustichini (2000) put forward to shed light on their findings. Let us call it the norm-switching theory.[39]

[39]The counterintuitive feature of the norm-switching theory is that fines are conventionally viewed as signaling inappropriateness rather than appropriateness. For example, the literature on the expressive function of law is based on the view that (in comparison to no law or punishment) a law supported by mild punishment creates deterrence by reinforcing the impression that the action is wrong; see, e.g., Sunstein (1996), Kahan (1997), and Funk (2007). Our (testable) hypothesis is that a law has a useful symbolic function when it is not otherwise obvious that an action is wrong. When that is already obvious, symbolic punishments will be counterproductive.

To learn more about why the fine backfires, we conducted an experiment to elicit social and personal appropriateness ratings, for picking up either 15 or 30 minutes late, and facing either no fine, a small fine of 1 dollar, or a big fine of 10 dollars (see Online Appendix for Instructions). Figure 7 displays our key findings.

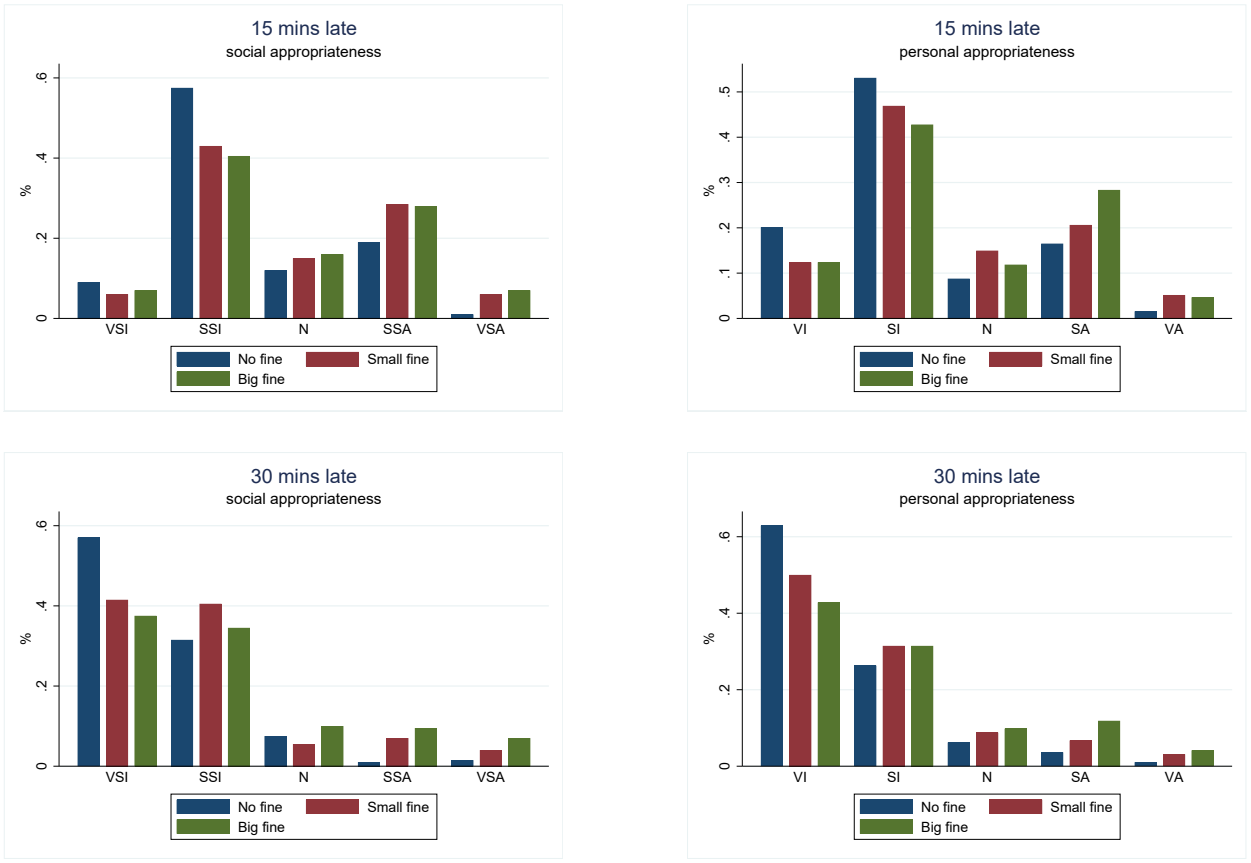


Figure 7. Daycare experiment: social (right) and personal (left) appropriateness ratings of being 15 minutes (top) or 30 minutes (bottom) late

As predicted both by the restitution hypothesis and the norm-switching hypothesis, collecting kids late is considered proscribed. Specifically, social ratings and personal ratings both yield more than 90 percent agreement that it is somewhat or very inappropriate to come late. This is significantly different from 1/2 with practical certainty ($p<$ 0.0001 for a test proportion, pre-registered). Moreover, it is considered more inappropriate when there is no fine than when fines are positive. For example, the average social appropriateness of being 15 minutes late is around -0.28 without a fine, but -0.07 with a small fine, and -0.06 with a large fine. All pairwise differences in social and personal appropriateness rating between no fine and positive fines are statistically significant at the $p < 0.001$ level (t-test, pre-registered), whereas not all differences between behavior under small and large fines are statistically significant (See Online Appendix for details). The first observation supports both the restitution and the norm-switching hypothesis, whereas the second fails to show decisive support for the restitution hypothesis over the

a	x_d, x_o	h	u_d
S	$-r, 0$	0	$-r$
N	$\kappa, -(k+m)$	$0, k+m$	$m - \delta(k+m)$

Table 7. Lost wallet experiment

norm-switching hypothesis.

4.7 Lost Wallet Experiment

In a large-scale field experiment conducted by Cohn et al (2019), the researchers turned in "lost" wallets at banks, hotels, post offices, and other public and private service providers around the world. The wallets took the form of transparent envelopes, which included a padlock key (in one treatment the key was removed), a grocery list in the local language, some business cards, and either no money or an amount of money corresponding to 13.5 US dollars, adjusted for local purchasing power. In a smaller fraction of cases, the money in the envelope amounted to 94 US dollars.

The business cards included the owner's email address, allowing the authors to measure the fraction of wallets that were honestly handled at the various places. A major finding is that a larger fraction of wallets is returned when they contain money. This finding was not predicted by experts, and it contradicts available models of unselfish behavior based on sympathy.

In the terminology of our model, the experiment can be described as follows. Decider finds Other's lost wallet. The wallet contains m units of money and a padlock key worth k to Other (the envelope, the few business cards, and the grocery shopping list presumably has little value to the owner). Returning the wallet costs r for Decider. This cost is the expected hassle of writing an email, engage in correspondence with the owner, and arranging for the return of the wallet. Suppose it is proscribed to keep the wallet. That is, only S is a just action. To simplify, we again set $\gamma = 0$. Thus, actions, material payoffs, harm, and utilities are as described in Table 7.

Comparing the utility from each of the two actions, we see that Decider returns Other's wallet (plays S) if $-r > m - \delta(k+m)$, i.e., if

$$\delta > \frac{r+m}{k+m}.$$

The first thing to note is that the condition is easier to satisfy when k is larger. Thus, the model is consistent with the finding that more envelopes are returned when they contain a padlock key than when they do not. It is also easily checked that the right-hand side is decreasing in the amount of money m when $r > k$.

Observation 7 *(i) The wallet is more likely to be returned if it contains a key. (ii) If $r > k$, wallets with more money are more likely to be returned.*

The intuition for the first result is obvious: The harm from keeping is greater when the envelope contains a key, and there is no benefit from keeping the key. The intuition for the second result is more subtle, since there is a benefit from keeping the extra money: If $r > k$, it takes substantial dutifulness δ to consider returning the envelope. Thus, for the people who are close to indifferent between keeping or returning the wallet, more money makes the guilt from keeping all of it grow faster than the profit from keeping the wallet.[40] We find it likely that many people consider their hassle cost to exceed the owner's value of the padlock key (padlocks often come with spare keys; alternatively, a padlock can be replaced at modest cost), and hence that the inequality is satisfied.[41]

Is it true that people consider that it is more inappropriate to keep a wallet with money in it than to keep the same wallet when it only contains a padlock key and other low-value items? To investigate this question, our experiment elicited social and personal appropriateness ratings for the lost-wallet situation.

As expected, there is virtual unanimity that it is inappropriate to keep the wallet, and in all conditions a large majority considers keeping to be *very* inappropriate. A drawback of this strong condemnation is that the floor effect makes it hard to test whether it is more inappropriate to keep a wallet with more money. In order to reduce the floor effect, we used seven appropriateness categories instead of five—with the category "inappropriate" inserted between "very inappropriate" and "somewhat inappropriate." These results are displayed in Figure 8, where "small money" denotes wallets with 13.5 dollars and "big money" denotes wallets with 94 dollars.

Comparing the social appropriateness of keeping a wallet with small money compared to one with no money, the average rating is about 0.05 lower (p=0.013). The same comparison between big money and small money yields a difference of about 0.04 (p=0.009). However, the personal rating differences are only about half as large, and not significant (p=0.145, p=0.096). (All comparisons with pre-registered t-tests.) Of course, these are rather weak tests of the model. A proper test of Observation 7 (ii) would require behavioral data with exogenous variation in the parameters r or k.

Let us briefly remark on the vast cross-country differences discovered by Cohn et al (2019) and Tannenbaum et al (2020). The authors document that for wallets containing money the propensity to contact the wallet's owner is around 80 percent in the Nordic countries, but only about 60 percent in the US and 20 percent in China. In our model, such differences can be explained by cross-country variation in the average level of the

[40] Of course, it is possible to return the wallet while keeping a fraction of the money. That very rarely occurs in the experiment, and the model suggests a reason: This action is dominated, since it fails to save r and it involves positive guilt.

[41] A testable implication of the model is that the presence of money would not have the same effect on the return-ratio if the wallet contained an item that is highly valuable to the owner but not to the finder.

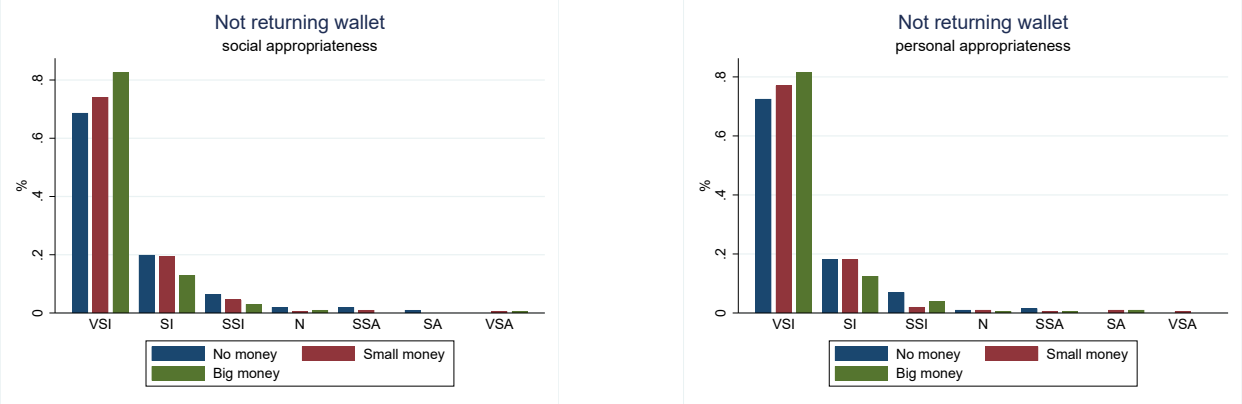


Figure 8. Lost waller experiment: social (right) and personal (left) appropriateness ratings of not returning

parameter δ. At first sight, this explanation may seem counter-intuitive, since the Nordic countries are among the least collectivistic countries in the world.[42] Thus, duty should matter less there, not more. However, a simple resolution to this puzzle is that high individualism goes together with high universalism. People in the Nordic countries experience moral duties toward everyone, rather than merely toward family, close relations, and other in-group members. Accordingly, Tannenbaum et al. (2020) find that a larger fraction of wallets are returned from countries with highly "generalized" or "universalistic" morality. (This is where the indicator-function mentioned at the beginning of Section 2 becomes important.) In short, we sympathize with the long-standing notion that the returning of lost wallets might provide a good measure of "social capital",[43] and we surmise that such behavior is most accurately ascribed to universalistic dutifulness.

5 Final Remarks

Levitt and List (2007) consider the role that experimentation plays in the natural and social sciences. They argue that social scientists face at least five exclusive challenges: 1) the presence of moral and ethical considerations; 2) the nature and extent of scrutiny of one's actions by others; 3) the context in which the decision is embedded; 4) self-selection of the individuals making the decisions; and 5) the stakes of the game. The presence of such considerations does not imply that social scientists should give up experimentation, but that there is an important and challenging role for theory in interpreting and extrapolating from the results.

Our model of social duties addresses items 1, 3, and 4. It is encouraging to us that accounting for entitlements offers a parsimonious explanation for several experimental

[42] See Inglehart (2018, Chapter 3) for a discussion of the world-wide distribution of the family of values that are defined as individualism, autonomy, and self-expression.

[43] See, e.g., Knack and Keefer, 1997.

regularities that most existing approaches fail to rationalize. Perhaps taking entitlements into account might overcome the weak external validity of laboratory experiments on the domain of unselfish behavior?[44] A way to design more externally valid laboratory experiments might be: First measure entitlements for the intended *field* application. Then construct a laboratory experiment in a way that reliably reflects these entitlements. In the final step, consider how potential policy-reforms affect behavior in the lab.

Our analysis has many limitations. Three stand out. First, we make no attempt to account for the effect of stakes (Levitt and List's item 5). Second, we only apply the model to understand simple situations in which a single person makes a decision. A next step is to consider interactions between several decision makers, and especially to deal with reciprocity. This is the topic of a planned companion paper. A third avenue for future research is to evaluate the quantitative aspects of the model. To what extent are the person-specific parameters robust across situations? What is the fraction of unselfish behavior that can be ascribed to the duty motive in comparison with other motives? To answer these questions, we need data for the same subjects across a range of different situations. (We are currently conducting such an analysis on the basis of data from Bruhin, Fehr, and Schunk, 2019.)

There are many other open questions as well. For all of our seven applications it is possible to think of experiments that would represent additional tests of our hypotheses. Theoretically, a natural avenue for future research is to investigate principles of entitlement that go beyond our current portable model, which depends entirely on configurations of feasible payoffs and the players' influence over these outcomes. For example, over what actions should we expect individuals to be granted freedom from duty? What are the roles of promises and contracting in establishing entitlements? How do social roles affect entitlements? Likewise, we do not offer a theory of when charity is salient and when it is not (the value of the parameter γ). Another limitation of our analysis is that we treat all duties as unwelcome. In reality, people often praise someone whose behavior is better than average, and desire for praise—or even praiseworthiness—could encourage voluntary sacrifice.[45] This would be a natural generalization of the model.[46]

Our purpose here is to explain behavior. However, to the extent that dutifulness is malleable through upbringing, education, and other public policies, the model is also

[44] For a documentation of this phenomenon, see Galizzi and Navarro-Martinez (2019) and the references therein.

[45] Desire to be praiseworthy relates to the concept of supererogation in theology and philosophy, where, roughly speaking, a supererogatory action goes above and beyond the call of duty. For an introduction see Heyd (2019); Dreier (2004) discusses supererogation in terms that are close to ours. For an introduction to the psychological literature on moral praise, see Anderson, Crockett, and Pizarro (2020).

[46] For example, Decider might obtain positive utility from praiseworthiness when communal value exceeds the customary level. Indeed, Adam Smith (1759, Part II, Sect II, Chap I) writes "That seems blamable which falls short of that ordinary degree of proper beneficence which experience teaches us to expect of every body; and, on the contrary, that seems praise-worthy which goes beyond it." For a closely related assumption, see Bénabou and Tirole (2011).

relevant for normative analysis. On one hand, raising dutifulness reduces the tension between common goals and individual desires. On the other hand, duties are constraining. The constraining nature of duties is the reason why liberal thinkers and politicians spent much of the 19th century trying to reduce social pressures on the individual. Suppose one takes a welfarist approach to normative analysis. Then, the social goal is to maximize a (weighted) sum of utilities. Does our utility function strike an acceptable balance between the benefits and costs of raising people's dutifulness? The benefits of increased dutifulness seem relatively uncontroversial. They consist of improvement of others' material outcomes. The costs are more problematic. Is it reasonable to include the full cost of compliance with duties as well as the guilt associated with non-compliance in the overall welfare measure? This is the approach taken by DellaVigna, List, and Malmendier (2012) and Andreoni, Rao, and Trachtman (2017).

This normative issue touches a sensitive nerve in moral philosophy. At one end of the spectrum, Nietzsche (1887) would argue that these costs are understated, as the imposition of guilt hampers the individual's natural freedom ("will to power"). At the other end of the spectrum, Hume (1751, Section IX, Part II, final paragraph) argues that the costs are exaggerated. According to Hume, the fulfillment of social duties brings a "peaceful reflection on one's own conduct" which is incomparably more worth than "the feverish, empty amusements of luxury and expense". That is, in Hume's view, lack of dutifulness represents either a failure of comprehension or excessive discounting. Hence, adapting our framework to conduct a normative analysis of moral instruction is another challenging task.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for Truth-Telling, *Econometrica* 87(4), 1115-1153.

Adams S.J. (1965) Inequity in Social Exchange, *Advances in Experimental Social Psychology* 2, 267-299.

Akerlof, G. and Kranton, R. (2000). Economics and Identity, *The Quarterly Journal of Economics* 115(3), 715-753.

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

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

Bartling, B., Engl, F., and Weber, R.A. (2014). Does Willful Ignorance Deflect Punishment? - An Experimental Study, *European Economic Review* 70, 512-524.

Bašić, Z., and Verrina, E. (2021). Personal norms—and not only social norms—shape economic behavior. MPI Collective Goods Discussion Paper, (2020/25).

Becker, G.S. (1974). A Theory of Social Interactions, *Journal of Political Economy* 82, 1063-1093.

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

Bénabou R. and Tirole J. (2011). Laws and Norms, NBER Working Paper 17579, November 2011.

Bernheim, B.D. (1994). A Theory of Conformity, *Journal of Political Economy* 102(5), 841-877.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge, MA: Cambridge University Press.

Binmore, K. (2005). *Natural Justice,* Oxford: Oxford University Press.

Blais, A. and Galais, C. (2016). Measuring the Civic Duty to Vote: A Proposal, *Electoral Studies* 41(1), 60-69.

Bolton G.E. and Ockenfels A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition, *American Economic Review* 90: 166-193.

Brekke, K.A., Kverndokk, S., and Nyborg, K. (2003). An Economic Model of Moral Motivation, *Journal of Public Economics* 9-10, 1967-1983.

Broberg, T., Ellingsen, T., and Johannesson, M. (2007). Is Generosity Involuntary? *Economics Letters* 94, 32-37.

Bruhin, A., Fehr, E., and Schunk, D. (2019). The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences, *Journal of the European Economic Association* 17(4),1025-1069.

Camerer, C.F. and Thaler, R.H. (1995). Anomalies: Ultimatums, Dictators and Manners, *Journal of Economic Perspectives* 9(2), 209-219.

Cappelen, A.W., Hole, A.D., Sørensen, E.Ø., and Tungodden, B. (2007). The Pluralism of Fairness Ideals: An Experimental Approach, *American Economic Review* 97(3), 818-827.

Cappelen, A.W., Konow, J., Sørensen, E.Ø., and Tungodden, B. (2013a) Just luck: An Experimental Study of Risk Taking and Fairness, *American Economic Review*103(3), 1398-1413.

Cappelen, A.W., Nielsen, U.H., Sørensen, E.Ø., Tungodden, B., and Tyran, J.-R. (2013b). Give and Take in Dictator Games, *Economics Letters* 118(2), 280-283.

Casson, M. (1991). *The Economics of Business Culture: Game Theory, Transaction Costs, and Economic Performance,* Oxford: Clarendon Press.

Charness, G. and Dufwenberg, M. (2006). Promises and Partnership, *Econometrica* 74, 1579-1601.

Charness G. and Rabin M. (2002). Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117: 817-869.

Cicero, M. Tullius (44 BC) *De Officiis (On Duty)*. English Translation: Walter Miller, Cambridge MA: Harvard University Press, 1913.

Cohn, A., Maréchal, M.A., Tannenbaum, D., and Zünd, C.L. (2019). Civic Honesty around the Globe, *Science* 365 (6448), 70-73.

Cooter, R. (1998). Expressive Law and Economics, *Journal of Legal Studies* 27 (Supplement 2), 585-607.

Cox, J.C., List, J.A., Price, M., Sadiraj, V., and Samek, A. (2019). Moral Costs and Rational Choice: Theory and Experimental Evidence, manuscript, Georgia State University.

Crawford, S.E.S. and Ostrom, E. (1995). A Grammar of Institutions, *American Political Science Review* 89(3), 582-600.

Curry, O.S., Mullins, D.A., and Whitehouse, H. (2019): Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies, *Current Anthropology* 60(1), 47-69.

Dana, J., Weber, R.A., and Kuang, J.X. (2007). Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness. *Economic Theory* 33: 67-80.

Dana, J, Cain, D.M., and Dawes, R.M. (2006). What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games, *Organizational Behavior and Human Decision Processes* 100(2), 193-201.

DellaVigna, S., List, J. A., and Malmendier U. (2012). Testing for Altruism and Social Pressure in Charitable Giving, *Quarterly Journal of Economics* 127(1), 1-56.

Dillenberger, D. and Sadowski, P. (2012). Ashamed to Be Selfish, *Theoretical Economics* 7(1), 99-124.

Dreber, A., Ellingsen, T., Johannesson, M., and Rand, D. (2013). Do People Care about Social Context? Framing Effects in Dictator Games, *Experimental Economics* 16, 349-371.

Dreier, J. (2004). Why Ethical Satisficing Makes Sense and Rational Satisficing Doesn't, In Michael Byron (ed.), *Satisficing and Maximizing*, Cambridge: Cambridge University Press, 131–154.

Duesenberry, J. (1960). Comment on "An Economic Analysis of Fertility" in *Demographic and Economic Change in Developed Countries,* edited by the Universities – National Bureau Committee for Economic Research, Princeton NJ: Princeton University Press.

Durkheim, E. (1958/1900). *Professional Ethics and Civil Morals,* Glencoe IL: Free Press. (Note: The manuscript was completed around 1900, but was first published in French in 1950.)

Edgeworth, F.Y. (1881). *Mathematical Psychics*, London: Kegan Paul.

Ellingsen T. and Johannesson M. (2008). Anticipated Verbal Feedback Induces Prosocial Behavior, *Evolution and Human Behavior* 29(2), 100-105.

Ellingsen T. and Johannesson M. (2008b). Pride and Prejudice: The Human Side of Incentive Theory, *American Economic Review* 98(3): 990-1008.

Enke, B., Rodríguez-Padilla, R., and Zimmermann, F. (2020). Moral Universalism and the Structure of Ideology, NBER Working Paper 27511.

Evren, Ö, and Minardi, S. (2017). Warm-glow Giving and Freedom to be Selfish, *Economic Journal* 127(603), 1381-1409.

Exley, C.L. and Petrie, R. The Impact of a Surprise Donation Ask, *Journal of Public Economics* 158, 152-167.

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

Feiler L. (2014). Patterns of Information Avoidance in Binary Choice Dictator Games, *Journal of Economic Psychology* 45, 253-267.

Freddi, E. (2021). Do People Avoid Morally Relevant Information? Evidence from the Refugee Crisis, *Review of Economics and Statistics* 103(4), 605-620.

Funk, P. (2007). Is There an Expressive Function of Law? An Empirical Analysis of Voting Laws with Symbolic Fines, *American Law and Economics Review* 9(1), 135-159.

Gächter, S. and Riedl, A. (2005). Moral Property Rights in Bargaining with Infeasible Claims, *Management Science* 51(2), 249-263.

Gächter, S. and Riedl, A. (2006). Dividing Justly in Bargaining Problems with Claims: Normative Judgments and Actual Negotiations, *Social Choice and Welfare* 27, 571-594.

Galizzi, M. and Navarro-Martinez, D. (2019). On the External Validity of Social Preference Games: A Systematic Lab-Field Study, *Management Science* 65(3), 976-1002.

Gelfand, M.J. et al (2011). Differences Between Tight and Loose Cultures: A 33-Nation Study, *Science* 332 (27 May), 1100-1104.

Gilligan, C. (1982). *In a Differerent Voice: Psychological Theory and Women's Development*, Cambridge MA: Harvard University Press.

Gneezy, U. and Rustichini, A. (2000). A Fine is a Price, *Journal of Legal Studies*, 29, 1-17.

Granovetter, M. (1985). Economic Action and Social Structure: The Problem of Embeddedness, *American Journal of Sociology* 91(3), 481-510.

Grossman, Z. (2014). Strategic Ignorance and the Robustness of Social Preferences, *Management Science* 60(11), 2659-2665.

Grossman, Z. and Van der Weele, J. (2017). Self-Image and Strategic Ignorance in Moral Dilemmas, *Journal of European Economic Association* 15(1), 171-217.

Haidt, J. (2008). Morality, *Perspectives on Psychological Science* 3(1), 65-72.

Henrich, J. et al (Eds.) (2004). *Foundations of Human Sociality*, Oxford: Oxford University Press.

Heyd, D. (2019). Supererogation, *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2019/entries/supererogation/

Huck, S., Kübler, D, Weibull, J. (2012). Social Norms and Economic Incentives in Firms, *Journal of Economic Behavior and Organization* 83(2), July 2012, 173-185.

Hume, D. (1751). *An Enquiry Concerning the Principles of Morals*, London: A Millar.

Hume, D. (1739-40). *A Treatise of Human Nature*. Norton, D. F., and Norton, M. J. (Eds.). (200). David Hume: A Treatise of Human Nature. OUP Oxford.

Inglehart, R. (2018). *Cultural Evolution: People's Motivations are Changing, and Reshaping the World*, Oxford: Oxford University Press.

Janoff-Bulman, R., Sheikh, S., Hepp, S. (2009). Proscriptive versus Prescriptive Morality: Two Faces of Moral Regulation, *Journal of Personality and Social Psychology* 96(3), 521-537.

Janoff-Bulman, R. and Carnes, N.C. (2013). Surveying the Moral Landscape: Moral Motives and Group-Based Moralities, *Personality and Social Psychology Review* 17(3), 219-236.

Jasso, G. and Opp, K.-D. (1997). Probing the Character of Norms: A Factorial Survey Analysis of the Norms of Political Action, *American Sociological Review* 62(6), 947-964.

Kahan, D.M. (1997). Social Influence, Social Meaning, and Deterrence, *Vanderbilt Law Review* 83, 349-395.

Kahan, D.M. (2013). Ideology, Motivated Reasoning, and Cognitive Reflection, *Judgment and Decision Making* 8(4), 407-424.

Kameda, T., Takezawa, M., Tindale, R.S., and Smith, C. (2002). Social Sharing and Risk Reduction: Exploring a Computational Algorithm for the Psychology of Windfall Gains, *Evolution and Human Behavior* 23, 11-33.

Kandel, E. and Lazear, E.P. (1992). Peer Pressure and Partnerships, *Journal of Political Economy* 100(4), 801-817.

Kandori, M. (1992). Social Norms and Community Enforcement, *Review of Economic Studies* 59(1), 63-80.

Kelley, H.H. and Thibaut, J.W. (1978). *Interpersonal Relations: A Theory of Interdependence*. New York, NY: Wiley.

Kimbrough, E.O. and Vostroknutov, A. (2016). Norms Make Preferences Social, *Journal of the European Economic Association* 14(3), 608-638.

Knack, S. and Keefer, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation, *Quarterly Journal of Economics* 112(4), 1251-1288.

Konow, J. (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions, *American Economic Review* 90(4), 1072--1091.

Konow, J. (1996). A Positive Theory of Economic Fairness, *Journal of Economic Behavior and Organization* 31(1), 13-35.

Konow, J. (2003). Which Is the Fairest One of All? A Positive Analysis of Justice Theories, *Journal of Economic Literature* 41(4), 1188-1239.

Korenok, O., Millner, E.L., and Razzolini, L. (2014). Taking, Giving, and Impure Altruism in Dictator Games, *Experimental Economics* 17(3), 488-500.

Krupka, E.L. and Weber, R.A. (2008). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? Institute for the Study of Labor, IZA DP No. 3860.

Krupka, E.L. and Weber, R.A. (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of European Economic Association* 11(3), 495-524.

Lazear, E.P., Malmendier, U., and Weber, R.A. (2012). Sorting, Prices, and Social Preferences, *American Economic Journal: Applied Economics* 4(1), 136-163.

Levitt, S. and List, J.A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *Journal of Economic Perspectives* 21, 153-174.

Lindbeck, A., Nyberg, S., and Weibull, J. (1999). Social Norms and Economic Incentives in the Welfare State, *Quarterly Journal of Economics* 114 (1), 1-35.

List, J.A. (2007). On the Interpretation of Giving in Dictator Games, *Journal of Political Economy* 115(3), 482-493.

López-Pérez, R. (2008). Aversion to Norm-breaking: A Model, *Games and Economic Behavior* 64, 237-267.

Malmendier, U., te Velde, V., and Weber, R.A. (2014). Rethinking Reciprocity, *Annual Review of Economics* 6, 849-874.

Nietzsche, F. (2003/1887). *On the Genealogy of Morals: A Polemic,* translated by H.B. Samuel, New York: Courier Dover Publications.

Nyborg, K. (2000). Homo Economicus and Homo Politicus: Interpretation and Aggregation of Environmental Values, *Journal of Economic Behavior and Organization* 42, 305-322.

Ostrom, E. (2000). Collective Action and the Evolution of Social Norms, *Journal of Economic Perspectives* 14(3),137-58.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2021). Data Quality of Platforms and Panels for Online Behavioral Research, *Behavior Research Methods* published online at https://doi.org/10.3758/s13428-021-01694-3.

Rabin, M. (1994). Cognitive Dissonance and Social Change, *Journal of Economic Behavior and Organization*, 23, 177-194.

Rabin, M. (1995). Moral Preferences, Moral Constraints, and Self-Serving Biases, Working Paper No 95-241, Department of Economics, University of California, Berkley.

Riker, W.H. and Ordeshook, P.C. (1968). A Theory of the Calculus of Voting, *American Political Science Review* 62(1), 25-42.

Ross, L. and Nisbett, R.E. (1991). *The Person and the Situation* New York: McGraw-Hill.

Rusbult, C.E. and Van Lange, P.A. (2008). Why We Need Interdependence Theory. *Social and Personality Psychology Compass*, 2(5), 2049-2070.

Saito, K. (2015). Impure Altruism and Impure Selfishness, *Journal of Economic Theory* 158(A), 336-70.

Sen, A. (1983). Liberty and Social Choice, *Journal of Philosophy* 80(1) 5-28.

Sen, A. (1993). Internal Consistency of Choice, *Econometrica* 61(3), 495-521.

Shafir, E. and Tversky, A. (1992). Thinking through Uncertainty: Nonconsequential Reasoning and Choice, *Cognitive Psychology* 24(4), 449-474.

Smith, A. (1759). *A Theory of Moral Sentiments*, printed for Andrew Millar, in the Strand; and Alexander Kincaid and J. Bell, in Edinburgh.

Stolte, J.F. (1987). The Formation of Justice Norms, *American Sociological Review* 12(1), 774-784.

Sobel, J. (2005). Interdependent Preferences and Reciprocity, *Journal of Economic Literature* 43(2), 392-436.

Sugaya, T. and Wolitzky, A. (2021). Communication and Community Enforcement, *Journal of Political Economy* 129(9), 2595-2628.

Sugden, R. (1986). *The Economics of Rights, Co-operation and Welfare*, Oxford: Basil Blackwell.

Sunstein, C. (1996). On the Expressive Function of Law, *University of Pennsylvania Law Review* 144, 2021-2053.

Tannenbaum, D., Cohn, A., Zünd, C.L., and Maréchal, M.A. (2020). What Do Cross-country Surveys Tell Us About Social Capital? University of Zürich Working Paper No. 352.

Tomasello M. (2020). The Moral Psychology of Obligation, *Behavioral and Brain Sciences* 43, e56: 1-58.

Umbeck, J (1977). The California Gold Rush; A Study of Emerging Property Rights, *Explorations in Economic History* 14(3), 197-226.

Weber, M. (1930/1905). *The Protestant Ethic and the Spirit of Capitalism*, London: Allen & Unwin.

Zerbe, R.O. Jr. and Anderson, C. Leigh (2001). Culture and Fairness in the Development of Institutions in the California Gold Fields, *Journal of Economic History* 61(1), 114-142.

Online Appendix to "A Model of Social Duties"

1 Proofs

1.1 Standard DE

In the paper's body, we followed the conventional practice of viewing the action a as the amount given, g. We here instead view the action as the amount kept, k. Since $k = 10 - g$ by definition, the translation between the two cases is immediate.

Proposition A1 (Observation 1) *In the standard dictator experiment with entitlement $e_o = 5\beta$ the Decider's action is*

$$k = \begin{cases} 10 & \text{if} \quad \delta < \dfrac{1}{1+2\alpha\gamma}; \\ 10 - 5\beta & \text{if} \quad \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma} \right); \\ 5 & \text{if} \quad \delta > \dfrac{1}{2\alpha\gamma}. \end{cases}$$

Proof. Utility is

$$u_d(k) = k - \delta \left(\max\{0, 5\beta - 10 + k\} + 2\alpha\gamma |k - 5| \right).$$

Note that $5\beta - 10 + k > 0$ implies $k > 5$. Thus $u_d(k)$ has three linear segments. If $k < 5$ then $u_d(k)$ is increasing in k. If $k \in (5, 10 - 5\beta)$ then

$$u_d(k) = k - 2\delta\alpha\gamma(k - 5),$$

which is increasing in k iff $\delta < 1/2\alpha\gamma$. If $k > 10 - 5\beta$ then

$$u_d(k) = k - \delta(5\beta - 10 + k + 2\alpha\gamma(k - 5)),$$

which is increasing in k iff $\delta < 1/(1 + 2\alpha\gamma)$. ■

1.2 DE with Taking-Option

Proposition A2 (Observation 2) *Consider a general Dictator experiment where Decider chooses $k \in [0, 10 + T]$, with $T \in \{0, 2, 10\}$, i.e. $k = 0$ corresponds to giving 10, $k = 10$ corresponds to neither giving nor taking, $k = 20$ corresponds to taking 10, and $k = 12$ corresponds to taking 2.*

The Decider's action is

$$k = \begin{cases} 10+T & \text{if } \delta < \dfrac{1}{1+2\alpha\gamma}; \\ 20-15\beta & \text{if } \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right), T = 10 \text{ and } e_o = e_o^{T=10} = 15\beta - 10; \\ 12-7\beta & \text{if } \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right), T = 2 \text{ and } e_o = e_o^{T=2} = 7\beta - 2; \\ 10-5\beta & \text{if } \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right), T = 0 \text{ and } e_o = e_o^{T=0} = 5\beta; \\ 10 & \text{if } \delta \in \left(\dfrac{1}{1+2\alpha\gamma}, \dfrac{1}{2\alpha\gamma}\right) \text{ and } e_o = 0; \\ 5 & \text{if } \delta > \dfrac{1}{2\alpha\gamma}. \end{cases}$$

Proof. Since Decider starts with 10 and Other starts with 0, we have $x_d(k) = k$ and $x_o(k) = 10 - a$. The entitlement of the Other is e_o. If it is endogenously determined then

$$\begin{aligned} e_o &= e_o^{T=10} = 5\beta - 10(1-\beta) = 15\beta - 10, \\ e_o &= e_o^{T=2} = 5\beta - 2(1-\beta) = 7\beta - 2 \\ e_o &= e_o^{T=0} = 5\beta. \end{aligned}$$

Note that, since $\beta \leq 1$ we have $e_o \leq 5$ in both cases. Alternatively we may want to set

$$e_0 = e_o^{T=10} = e_o^{T=2} = 0.$$

Utility is

$$u_d(k) = k - \delta\left(\max\{0, e_o - 10 + k\} + 2\alpha\gamma|k-5|\right).$$

Note that $\max\{0, e_o - 10 + k\} > 0$ implies $k > 5$. Thus, $u_d(k)$ has three linear segments, as follows. If $k < 5$ then $u_d(k)$ is increasing in k. If $k \in (5, 10-5\beta)$ then

$$u_d(k) = k - 2\delta\alpha\gamma(k-5),$$

which is increasing in k iff $\delta < 1/2\alpha\gamma$. If $k > 10 - 5\beta$ then

$$u_d(k) = k - \delta(5\beta - 10 + k + 2\alpha\gamma(k-5)),$$

which is increasing in k iff $\delta < 1/(1+2\alpha\gamma)$. Optimum is at $k = 5$, $k = 10 - e_o$ or

$k = 10 + T$. Thus,

$$k = \begin{cases} 10 + T & \text{if} \quad \delta < \dfrac{1}{1 + 2\alpha\gamma}; \\ 10 - e_o & \text{if} \quad \delta \in \left(\dfrac{1}{1 + 2\alpha\gamma}, \dfrac{1}{2\alpha\gamma} \right); \\ 5 & \text{if} \quad \delta > \dfrac{1}{2\alpha\gamma}. \end{cases}$$

■

The behavioral data of List and others show that the share of giving decreases as we move from the standard DE ($T = 0$) to the taking versions ($T = 2, 10$) . Similarly, appropriateness ratings show that it becomes appropriate to leave less to the Other. This is reflected in our model: as T increases the derived entitlement e_o^T decreases for agents with $\beta < 1$. Thus the Decider can leave less to the Other without doing something inappropriate. Moreover, we believe that adding the taking option suggests to subjects that the entitlement is at 0. Again, this can explain the shift in behavior and appropriateness ratings. In particular it explains why the appropriateness ratings for the taking treatments display a sharp decrease in appropriateness at zero.

The fraction of equal splits drops dramatically when taking options are added. In List's experiment 25% chose $k = 5$ in the standard DE compared to 9% when $T = 2$ and 6% when $T = 10$. Our model can explain this by assuming that there is a fraction of agents with $\beta = 1$. These people perceive the entitlement of Other to be 5 in the standard DE. Once the taking option is introduced they perceive the entitlement to be zero.

1.3 DE with Exit Option

Proposition A3 (Observation 3) *Suppose exiting carries no blame. D sets* $10 - 5\beta$ *and remains if*

$$\frac{1}{1 + 2\alpha\gamma} < \delta < \frac{1 - 5\beta}{10\alpha\gamma\,(1 - \beta)} < \frac{1}{2\alpha\gamma}.$$

D sets $k = 10$ *and remains if*

$$\delta < \min \left\{ \frac{1}{5\,(\beta + 2\alpha\gamma)}, \frac{1}{1 + 2\alpha\gamma} \right\}.$$

Otherwise D exits.

Remark A1 *The condition*

$$\delta < \frac{1 - 5\beta}{10\alpha\gamma\,(1 - \beta)} \iff \beta < \frac{1 - \delta 10\alpha\gamma}{5 - \delta 10\alpha\gamma},$$

requires $\beta < 1/5$, *which implies* $k = 10 - 5\beta > 9$. *Thus D only sticks with her initial choice if*

it was at least 9. Moreover, the condition

$$\frac{1}{1+2\alpha\gamma} < \frac{1-5\beta}{10\alpha\gamma\left(1-\beta\right)},$$

is equivalent to $8\alpha\gamma + 5\beta < 1$. *Thus if* $8\alpha\gamma + 5\beta > 1$ *then all who set* $k < 10$ *exit.*

Proof. D sets $k = 5$ if $\delta > 1/\left(2\alpha\gamma\right)$, but does not stick with this choice since

$$u_d\left(5\right) = 5 < 9 = u_d\left(Exit\right).$$

D sets $k = 10$ if $\delta < 1/\left(1+2\alpha\gamma\right)$, obtaining utility $u_d\left(10\right) = 10 - \delta\left(5\beta + 10\alpha\gamma\right)$, and sticks with this choice if

$$10 - 5\delta\left(\beta + 2\alpha\gamma\right) > 9 \Longleftrightarrow \delta < \frac{1}{5\left(\beta + 2\alpha\gamma\right)}.$$

Thus, D sets $k = 10$ and sticks with it if

$$\delta < \min\left\{\frac{1}{5\left(\beta+2\alpha\gamma\right)}, \frac{1}{1+2\alpha\gamma}\right\}.$$

Note

$$\frac{1}{5\left(\beta+2\alpha\gamma\right)} < \frac{1}{1+2\alpha\gamma} \Longleftrightarrow 8\alpha\gamma + 5\beta > 1.$$

D sets $k = 10 - 5\beta$ if

$$\delta \in \left(\frac{1}{1+2\alpha\gamma}, \frac{1}{2\alpha\gamma}\right),$$

thereby obtaining utility $u_d\left(10\right) = 10 - 10\delta\left(\alpha\gamma\left(1-\beta\right)\right)$, and sticks with this choice if

$$10 - 5\beta - 10\delta\alpha\gamma\left(1-\beta\right) > 9 \Longleftrightarrow \delta < \frac{1-5\beta}{10\alpha\gamma\left(1-\beta\right)}.$$

Thus, D sets $k = 10 - 5\beta$ and sticks with it if

$$\frac{1}{1+2\alpha\gamma} < \delta < \min\left\{\frac{1-5\beta}{10\alpha\gamma\left(1-\beta\right)}, \frac{1}{2\alpha\gamma}\right\}.$$

Note that

$$\frac{1}{1+2\alpha\gamma} < \frac{1-5\beta}{10\alpha\gamma\left(1-\beta\right)} \Longleftrightarrow 8\alpha\gamma + 5\beta < 1.$$

Hence, if $8\alpha\gamma + 5\beta > 1$ then all who chose $k = 10 - 5\beta$ exit, whereas if $8\alpha\gamma + 5\beta < 1$ then those with

$$\frac{1}{1+2\alpha\gamma} < \delta < \frac{1-5\beta}{10\alpha\gamma\left(1-\beta\right)} < \frac{1}{2\alpha\gamma}$$

chose $k = 10 - 5\beta$ and remain. ■

Proposition A4 (Observation 4) *Suppose exiting carries no blame from obligation (because presence of the exit option nullifies the entitlement of other), but still carries blame from responsibility. D sets $k = 5$ and remains if*

$$\delta > \frac{1}{2\alpha\gamma}.$$

D sets $k = 10 - 5\beta$ and remains if

$$\max\left\{\frac{5\beta - 1}{10\gamma\alpha\beta + \gamma(1-\alpha)}, \frac{1}{1+2\alpha\gamma}\right\} < \delta < \frac{1}{2\alpha\gamma}.$$

D sets $k = 10 - 5\beta$ and exits if

$$\frac{1}{1+2\alpha\gamma} < \delta < \frac{5\beta - 1}{10\gamma\alpha\beta + \gamma(1-\alpha)}.$$

D sets $k = 10$ and remains if

$$\delta < \min\left\{\frac{1-5\beta}{5\beta - (1-\alpha)\gamma}, \frac{1}{1+2\alpha\gamma}\right\}.$$

D sets $k = 10$ and exits if

$$\frac{1-5\beta}{5\beta - (1-\alpha)\gamma} < \delta < \frac{1}{1+2\alpha\gamma}.$$

Remark A2 *Note that*

$$\frac{1}{1+2\alpha\gamma} < \frac{5\beta - 1}{10\gamma\alpha\beta + \gamma(1-\alpha)} \iff (1+\alpha)\gamma < 5\beta - 1.$$

Proof. D sets $k = 5$ if $\delta > 1/2\alpha\gamma$ and remains if

$$\begin{aligned} u_d(E) &= 9 - \delta\gamma(10 - (9 - 9\alpha)) = 9 - \delta\gamma(1 + 9\alpha) < 5 = u_d(5). \\ &\iff \delta > \frac{4}{\gamma(1+9\alpha)}. \end{aligned}$$

Note that

$$\frac{4}{\gamma(1+9\alpha)} < \frac{1}{2\alpha\gamma},$$

implying that all who set $k = 5$ remain.

D sets $k = 10$ if $\delta < 1/(1+2\alpha\gamma)$ and remains if

$$\begin{aligned} u_d(k = 10 - 5\beta) &= 10 - 5\beta - \delta(5\beta + 10\alpha\gamma) > 9 - \delta\gamma(1+9\alpha) = u_d(E) \\ &\iff \delta < \frac{1-5\beta}{5\beta - (1-\alpha)\gamma}. \end{aligned}$$

Thus, D sets $k = 10$ and sticks with her choice if

$$\delta < \min\left\{\frac{1-5\beta}{5\beta-(1-\alpha)\gamma}, \frac{1}{1+2\alpha\gamma}\right\}.$$

D sets $k = 10-5\beta$ if

$$\delta \in \left(\frac{1}{1+2\alpha\gamma}, \frac{1}{2\alpha\gamma}\right),$$

and sticks with her choice if

$$\begin{aligned} u_d(k=10-5\beta) &= 10-5\beta-10\delta\gamma\alpha(1-\beta) > 9-\delta\gamma(1+9\alpha) = u_d(E) \\ &\iff \frac{5\beta-1}{10\gamma\alpha\beta+\gamma(1-\alpha)} < \delta. \end{aligned}$$

However, it can be verified that,

$$\frac{5\beta-1}{10\gamma\alpha\beta+\gamma(1-\alpha)} < \frac{1}{2\alpha\gamma}.$$

Thus, D sets $k = 10-5\beta$ and remains if

$$\max\left\{\frac{5\beta-1}{10\gamma\alpha\beta+\gamma(1-\alpha)}, \frac{1}{1+2\alpha\gamma}\right\} < \delta < \frac{1}{2\alpha\gamma},$$

whereas D sets $k = 10-5\beta$ and exits if

$$\frac{1}{1+2\alpha\gamma} < \delta < \frac{5\beta-1}{10\gamma\alpha\beta+\gamma(1-\alpha)}.$$

■

Proposition A5 *Suppose exiting violates an obligation. Then no one exits.*

Proof. In the standard DE the entitlement of the other is 5β. Hence, exit creates harm of 5β for the other. We have

$$\begin{aligned} u_d(k=9) &= 9-\delta(\max\{0,9-5\beta\}+\gamma(10-(10-\alpha|9-1|))) \\ &= 9-\delta(9-5\beta+8\alpha\gamma) \\ &> 9-\delta(9-5\beta+9\alpha\gamma+\gamma) \\ &= 9-\delta(\max\{0,9-5\beta\}+\gamma(10-(9-\alpha|9-0|))) \\ &= u_d(E), \end{aligned}$$

meaning that everyone prefers $k = 9$ in the original game to the exit option. ■

1.4 DE with Hidden Information

KW-elicitation indicates that a substantial portion of subjects find it inappropriate to chose option A in the non-aligned state (47% find it very socially inappropriate and 30% find it somewhat socially inappropriate). This leads us to assume $e_o = 5$. If entitlements are instead derived from might and right, then in the non-aligned state $e_o^* = 5\beta + (1-\beta) = 1 + 4\beta$. Thus, from (1c), Other's entitlement in the non-aligned state is

$$e_o = \arg\min_{x_o \in \{1,5\}} |x_o - (1+4\beta)| = \begin{cases} 1 & \text{if } \beta < 1/2; \\ 5 & \text{if } \beta > 1/2. \end{cases}$$

In the aligned state Others's entitlement is

$$e_o^* = e_o = 5\beta + 5(1-\beta) = 5.$$

Proposition A6 (Observation 5) *Decider prefers RBA if*

$$\delta > \frac{1}{\gamma(3+5\alpha)}.$$

If the inequality is reversed, Decider prefers NA (if $e_o = 5$ in the non-aligned state), or NA and RAA (if $e_o = 1$ in the non-aligned state).

Proof. Suppose $e_o = 5$ in the non-aligned state. The following table summarises the strategies available to Decider, and their consequences.

	Non-aligned			Aligned		
	x_d, x_o	h_o	s	x_d, x_o	h_o	s
RAA	$6,1$	4	$3+5\alpha$	$6,5$	0	0
RAB	$6,1$	4	$3+5\alpha$	$5,1$	4	$5+3\alpha$
RBA	$5,5$	0	0	$6,5$	0	0
RBB	$5,5$	0	0	$5,1$	4	$5+3\alpha$
NA	$6,1$	0	$3+5\alpha$	$6,5$	0	0
NB	$5,5$	0	0	$5,1$	0	$5+3\alpha$

For example, the shortage induced by RAB in the non-aligned state is $10-(7-5\alpha) = 3+5\alpha$ and the shortage induced by RAB in the aligned state is $11-\alpha-(6-4\alpha) = 5+3\alpha$.

Since states are equally probable, the expected utility is

$$\begin{aligned}
u_d(RAA) &= 6 - \delta\frac{1}{2}(4 + \gamma(3 + 5\alpha)); \\
u_d(RAB) &= 5.5 - 4\delta(1 + \gamma(1 + \alpha)); \\
u_d(RBA) &= 5.5; \\
u_d(RBB) &= 5 - \delta\frac{1}{2}(4 + \gamma(5 + 3\alpha)); \\
u_d(NA) &= 6 - \delta\frac{1}{2}(\gamma(3 + 5\alpha)); \\
u_d(NB) &= 5 - \delta\frac{1}{2}(\gamma(5 + 3\alpha)).
\end{aligned}$$

Note that RBA dominates NB and RBB, and that NA dominates RAA and RAB. Hence we only need to compare RBA and NA.

$$u_d(RBA) > u_d(NA) \iff \delta > \frac{1}{\gamma(3 + 5\alpha)}.$$

Suppose $e_o = 1$ in the non-aligned state. Harm is now zero in the non-aligned state. It is still the case that RBA dominates NB and RBB However, now NA and RAA yield the same payoff and they both dominate RAB. Still, the comparison between RBA and NA is the same

$$u_d(RBA) > u_d(NA) = u_d(RAA) \iff \delta > \frac{1}{\gamma(3 + 5\alpha)}.$$

■

2 Additional Theoretical Results and Proofs

2.1 DE with Taking-Option: Fixed Cost

Suppose that, in addition to the cost of blame, the decider suffers a fixed cost β^H when creating harm and a fixed cost β^S when creating shortage. The utility function is then

$$u_D(k) = k - \delta(h(k) + s(k)) - \beta^S \mathbb{I}_{\{s(k)>0\}} - \beta^H \mathbb{I}_{\{h(k)>0\}}.$$

Proposition A7 *Consider a general Dictator experiment where Decider chooses $k \in [0, 10+T]$, with $T \in \{0, 2, 10\}$, i.e. $k = 0$ corresponds to giving 10, $k = 0$ corresponds to neither giving nor taking, $k = 20$ corresponds to taking 10, and $k = 12$ corresponds to taking 2. Assume*

$$\begin{aligned} \alpha &= \beta = 1/2, \\ \gamma &= 1, \\ \beta^S &< 1. \end{aligned}$$

If $T = 0$ and $e_o = \frac{5}{2}$,the Decider's action is

$$k = \begin{cases} 10 & \text{if } \delta < \dfrac{1}{2} - \dfrac{\beta^H}{5}; \\ 10 - 5\beta & \text{if } \delta \in \left(\dfrac{1}{2} - \dfrac{\beta^H}{5}, 1 - \dfrac{2\mu^S}{5}\right); \\ 5 & \text{if } \delta > 1 - \dfrac{2\mu^S}{5}. \end{cases}$$

If $T = 2$ and $e_o = 0$, the Decider's action is

$$k = \begin{cases} 12 & \text{if } \delta < \dfrac{1}{2} - \dfrac{\beta^H}{4}; \\ 10 & \text{if } \delta \in \left(\dfrac{1}{2} - \dfrac{\beta^H}{4}, 1 - \dfrac{\beta^S}{5}\right); \\ 5 & \text{if } \delta > 1 - \dfrac{\beta^S}{5}. \end{cases}$$

If $T = 10$ and $e_o = 0$, the Decider's action is

$$k = \begin{cases} 20 & \text{if } \delta < \dfrac{1}{2} - \dfrac{\beta^H}{20}; \\ 10 & \text{if } \delta \in \left(\dfrac{1}{2} - \dfrac{\beta^H}{20}, 1 - \dfrac{\beta^S}{5}\right); \\ 5 & \text{if } \delta > 1 - \dfrac{\beta^S}{5}. \end{cases}$$

Remark A3 *Note that individuals with δ such that*

$$\frac{1}{2} - \frac{\beta^H}{4} < \delta < \frac{1}{2} - \frac{\beta^H}{20},$$

pick $k = 20$ when $T = 10$, but do not pick $k = 12$ when $T = 2$. This accounts for the fact that there are more people who pick $k = 20$ when $T = 10$, than $k = 12$ when $T = 2$.

Note that individuals with δ such that

$$1 - \frac{2\mu^S}{5} < \delta < 1 - \frac{\beta^S}{5},$$

pick $k = 5$ when $T = 0$, but do not pick $k = 5$ when $T = 2$ or $T = 10$. This helps explain why the fraction of equal splits drops dramatically when taking options are added. As mentioned, this can also be explained without fixed costs, by assuming that a fraction of the subjects have $\beta = 1$.

Proof. Utility is

$$\begin{aligned} u_D(k) &= k - \delta\left(\max\{0, e_o - 10 + k\} + |k - 5|\right) \\ &\quad -\beta^S \mathbb{I}_{\{s(k)>0\}} - \beta^H \mathbb{I}_{\{h(k)>0\}}. \end{aligned}$$

As before, $u_D(k)$ has three linear segments, with the optimum at $k = 5$, $k = 10 - e_o$ or $k = 10 + T$.

Suppose $T = 0$ and $e_o = \frac{5}{2}$. We have

$$\begin{aligned} u_D(5) &= 5; \\ u_D(10 - 5\beta) &= 10 - \frac{5}{2} - 5\delta\alpha\gamma - \beta^S; \\ u_D(10) &= 10 - \frac{15}{2}\delta - \beta^S - \beta^H, \end{aligned}$$

and

$$\begin{aligned} u_D(5) &> u_D(10 - 5\beta) \Longleftrightarrow \delta > 1 - \frac{2\mu^S}{5}; \\ u_D(5) &> u_D(10) \Longleftrightarrow \delta > \frac{2}{3} - \frac{2\mu^S}{15}; \\ u_D(10) &> u_D(10 - 5\beta) \Longleftrightarrow \delta < \frac{1}{2} - \frac{\beta^H}{5}. \end{aligned}$$

Note that if $\beta^S < 1$ then

$$\frac{1}{2} - \frac{\beta^H}{5} < \frac{2}{3} - \frac{2\mu^S}{15} < 1 - \frac{2\mu^S}{5}.$$

Suppose $T = 2$ and $e_o = 0$. We have

$$\begin{aligned} u_D(5) &= 5; \\ u_D(10) &= 10 - 5\delta - \beta^S; \\ u_D(12) &= 12 - 9\delta - \beta^S - \beta^H, \end{aligned}$$

and

$$\begin{aligned} u_D(5) &> u_D(10) \Longleftrightarrow \delta > 1 - \frac{\beta^S}{5}; \\ u_D(5) &> u_D(12) \Longleftrightarrow \delta > \frac{7}{9} - \frac{\beta^S + \beta^H}{9}; \\ u_D(12) &> u_D(10) \Longleftrightarrow \delta < \frac{1}{2} - \frac{\beta^H}{4}. \end{aligned}$$

Note that if $\beta^S \leq 1$ then

$$\frac{1}{2} - \frac{\beta^H}{4} < \frac{7}{9} - \frac{\beta^S + \beta^H}{9} < 1 - \frac{\beta^S}{5}.$$

Suppose $T = 10$ and $e_o = 0$. We have

$$\begin{aligned} u_D(5) &= 5; \\ u_D(10) &= 10 - 5\delta - \beta^S; \\ u_D(20) &= 20 - 25\delta - \beta^S - \beta^H, \end{aligned}$$

and

$$\begin{aligned} u_D(5) &> u_D(10) \Longleftrightarrow \delta > 1 - \frac{\beta^S}{5}; \\ u_D(5) &> u_D(20) \Longleftrightarrow \delta > \frac{3}{5} - \frac{\beta^S + \beta^H}{25}; \\ u_D(20) &> u_D(10) \Longleftrightarrow \delta < \frac{1}{2} - \frac{\beta^H}{20}. \end{aligned}$$

Note that if $\beta^S \leq 1$ then

$$\frac{1}{2} - \frac{\beta^H}{20} < \frac{3}{5} - \frac{\beta^S + \beta^H}{25} < 1 - \frac{\beta^S}{5}.$$

■

2.2 DE with Hidden Information: Quadratic Cost of Blame

In all examples so far, the decider eventually learns the outcome of the game. At the end of the game she knows exactly how much shortage and harm she has caused, and consequently she knows exactly how blameworthy she is. Ex post her utility at end node z is

$$u_D(z) = x_o(z) - \delta(h(z) + s(z)).$$

In this case it makes perfect sense to assume that the Deciders ex ante chooses and action that maximizes expected utility

$$\begin{aligned}\mathbb{E}[u_D(k)] &= \mathbb{E}_a[x_o(z)] - \delta\mathbb{E}_a[(h(z) + s(z))] \\ &= x_o(k) - \delta(\mathbb{E}_a[h(z)] + \mathbb{E}_a[s(z)]).\end{aligned}$$

Here, expectation is taken with respect to the probability distribution over end nodes induced by action k. Since the Decider always is informed about her own material payoff, which is a deterministic function of her action, we have $\mathbb{E}_a[x_o(z)] = x_o(a)$.

However, in the current game, if the Decider remains uninformed she is never informed about the state of the game. Thus, she will never know how much shortage and harm she has caused, since she does not know at what end node she is. In this case it makes sense to assume that she suffers from the expected harm and expected shortage she has induced. Ex post her utility at end node z is

$$u_D(z) = x_o(z) - \delta(\mathbb{E}_a[h(z)] + \mathbb{E}_a[s(z)]),$$

where expectation is taken with respect to the probability distribution over end nodes induced by action k. Ex ante she maximizes expected utility

$$\begin{aligned}\mathbb{E}[u_D(k)] &= \mathbb{E}_a[x_o(z)] - \delta\mathbb{E}_a[(\mathbb{E}_a[h(z)] + \mathbb{E}_a[s(z)])] \\ &= x_o(k) - \delta(\mathbb{E}_a[h(z)] + \mathbb{E}_a[s(z)]).\end{aligned}$$

Clearly, with a linear specification of blame and utility there is no difference between these viewpoints. But for non-linear specifications there is a difference.

In the main text we have worked with a simple linear specification that delivers the main results that we are after. More realistically we believe that the cost of blame may be convex, in addition to fixed costs for positive shortage and positive harm, as described before. For simplicity, suppose that utility is quadratic in blame. Ex post Utility at end node z is

$$u_D(z) = x_o(z) - \delta\left(\left(\mathbb{E}_a\left[h\left(z'\right)\right] + \mathbb{E}_a\left[s\left(z'\right)\right]\right)^2 - \beta^S\mathbb{I}_{\{\mathbb{E}_a[s(z')]>0\}} - \beta^H\mathbb{I}_{\{\mathbb{E}_a[h(z')]>0\}}\right),$$

where $\mathbb{E}_a[h(z')]$ and $\mathbb{E}_a[h(z')]$ represent expectations taken over the set of terminal nodes that cannot be distinguished from the actually reached end node z.

Proposition A8 *Let p be the probability of the aligned state. Suppose there is a fixed cost β^H of creating creating harm (but no fixed cost of creating shortage). If*

$$\delta > \delta^*(p) := \frac{1}{(1-p)(\gamma(3+5\alpha))^2},$$

then Decider prefers RBA. If the inequality is reversed the decider prefers NA. The threshold $\delta^*(p)$ is increasing in p.

Proof. Suppose $e_o = 5$ in the non-aligned state. The following table summarizes the strategies available to the Decider, and their consequences.

	Non-aligned			Aligned		
	x_d, x_o	h_o	s	x_d, x_o	h_o	s
RAA	$6,1$	4	$3+5\alpha$	$6,5$	0	0
RAB	$6,1$	4	$3+5\alpha$	$5,1$	4	$5+3\alpha$
RBA	$5,5$	0	0	$6,5$	0	0
RBB	$5,5$	0	0	$5,1$	4	$5+3\alpha$
NA	$6,1$	0	$3+5\alpha$	$6,5$	0	0
NB	$5,5$	0	0	$5,1$	0	$5+3\alpha$

Utility is

$$\begin{aligned}
u_d(RAA) &= 6-\delta(1-p)\left((4+\gamma(3+5\alpha))^2+\beta^H\right)\\
u_d(RAB) &= 6-p-\delta\left((1-p)(4+\gamma(3+5\alpha))^2+p(4+\gamma(5+3\alpha))^2+\beta^H\right)\\
u_d(RBA) &= 5+p\\
u_d(RBB) &= 5-\delta p\left((4+\gamma(5+3\alpha))^2+\beta^H\right)\\
u_d(NA) &= 6-\delta((1-p)\gamma(3+5\alpha))^2\\
u_d(NB) &= 5-\delta(p\gamma(5+3\alpha))^2.
\end{aligned}$$

Note that RBA dominates NB and RBB. Furthermore NA dominates RAA and NA dominates RAB. Thus, we only need to compare RBA and NA.

$$u_d(RBA) > u_d(NA) \iff \delta > \frac{1}{(1-p)(\gamma(3+5\alpha))^2}.$$

Suppose $e_o = 1$ in the non-aligned state. Harm is now zero in the non-aligned state. It is still the case that RBA dominates NB and RBB. Furthermore NA dominates RAA and

NA dominates RAB and the condition for $u_d(RBA) > u_d(NA)$ is the same as before. ■

3 Experimental Procedures

In our experiments subjects read descriptions of a number of situations. In each situation there is one person who makes a choice between different actions. The subjects are asked to rate the appropriateness of the different actions.

In some treatments the situations are different experimental games (including the dictator situations examined in this paper). In other treatments the situations consist of the lost wallet and day-care center situations.

In some treatments the subjects are asked to rate the social appropriateness of actions and in other treatments they are asked to rate the personal appropriateness of actions.

The subjects earn a fixed participation reward. In addition, subjects in the social appropriateness treatments earn a bonus payment if their answer on a randomly selected question matches the most common answer on that question.

3.1 Response Scales

In most treatments we use a 5-item scale. For the social appropriateness treatments the scale is: *Very Socially Inappropriate*, *Somewhat Socially Inappropriate*, *Neither Socially Appropriate nor Inappropriate*, *Somewhat Socially Appropriate*, *Very Socially Appropriate*. For the personal appropriateness treatments the scale is: *Very Inappropriate*, *Somewhat Inappropriate*, *Neither Socially Appropriate nor Inappropriate*, *Somewhat Appropriate*, *Very Appropriate*. In some treatments involving the lost wallet and day-care center situations we use as 7-item scale in order to mitigate floor/ceiling effects. For the social appropriateness treatments the scale is: *Very Socially Inappropriate*, *Socially Inappropriate, Somewhat Socially Inappropriate*, *Neither Socially Appropriate nor Inappropriate*, *Somewhat Socially Appropriate*, *Socially Appropriate, Very Socially Appropriate*. For the personal appropriateness treatments the scale is analogous but without the qualifier *Social*. We use the following abbreviations:

Response	Abbreviation
Very (Socially) Inappropriate	VI
(Socially) Inappropriate	I
Somewhat (Socially) Inappropriate	SI
Neither (Socially) Appropriate nor Inappropriate	N
Somewhat (Socially) Appropriate	SA
(Socially) Appropriate	A
Very (Socially) Appropriate	VA

We convert responses to numerical values where the least appropriate option (VI) is given the value -1, the most appropriate option (VA) is given the value 1, and remaining options are given values that create equal distances between any two adjacent options,

implying that the neutral option (N) has value 0. That is, for the 5-item scale we convert responses to numbers as follows: $VI = -1$, $SI = -1/2$, $N = 0$, $SA = 1/2$, $VA = 1$. For the 7-item scale set convert responses to numbers as follows: $VI = -1$, $I = -2/3$, $SI = -1/3$, $N = 0$, $SA = 1/3$, $A = 2/3$, $VA = 1$. With the responses converted to numerical values we can calculate average ratings. Let R_a denote the average rating of alternative/action a.

3.2 Dictator Experiments

Appropriateness elicitation in dictator games was conducted in treatments that also elicited appropriateness in a number of other experimental game situations, which we will analyze in a companion paper. In total the following game situations were examined: Standard dictator game (DG), DG with taking option, DG with exit option, DG with Information avoidance, DG with production, DG with quiz, Mini-UG, Ultimatum Game (UG), and Trust Game (TG). See the experimental instructions for a complete description of the situations as they were presented to subjects.

Each subject faced 3 or 4 of these situation (in randomized order) as follows:

1. Each subject faced one of the following:
   (a) DG standard
   (b) DG with exit option
   (c) DG with take-1 option
   (d) DG with take-5 option
2. Each subject faced one of the following:
   (a) Mini UG with 5:5 outside option
   (b) Mini UG with 2:8 outside option
   (c) UG Proposer
   (d) UG Responder (rating different cut-offs for rejection in one single question)
3. Each subject faced one of the following:
   (a) DG with information avoidance Baseline
   (b) DG with unformation avoidance Treatment
   (c) TG Proposer
   (d) TG Responder (rating shares sent back for different amounts received).

4. Subject who did not face either TG responder or DG with information avoidance Treatment faced one of the following:

   (a) DG with quiz

   (b) DG with production

We ran two experiments involving these experimental games, one eliciting social appropriateness and one eliciting personal appropriateness. In each of them there were 600 subjects, meaning that each game was encountered by 150 subjects under each mode of appropriateness. Subjects earned a participation reward of GBP 3. In the social appropriateness treatments subjects earned an additional GBP 4 in case their response to a randomly selected question was equal to the most frequently given response on that questions.

3.3 Lost Wallet and Day-Care Center Situations

Each subject faced two sets of situations: lost wallet situations and day-care center situations. Each set of situations consists of three very similar situations. Specifically, in the lost wallet situation the wallet contains (i) no money, (ii) small money, or (iii) big money, and the day-care center situations there is (i) no fine, (ii) a small fine, or (iii) a large fine. Each subject faces each of the three situations in a set in random order. Moreover, the order of the two sets of situations is randomized.

We ran four experiments involving the lost wallet situations and day-care center situations, varying the mode of appropriateness (social or personal) and the rating scale (5 or 7 items). In each of them there were 200 subjects. Subjects earned a participation reward of GBP 2. In the social appropriateness treatments subjects earned an additional GBP 2.5 in case their response to a randomly selected question was equal to the most frequently given response on that questions.

4 Main Hypotheses

In this section we describe all our pre-registered main hypotheses. The three preregistrations can be found at the following urls:
https://mfr.de-1.osf.io/render?url=https://osf.io/3q8ma/direct%26mode=render%26action=download%26mode=render,

https://mfr.de-1.osf.io/render?url=https://osf.io/yjksa/?direct%26mode=render%26action=download%26mode=render,

https://mfr.de-1.osf.io/render?url=https://osf.io/qzpfk/?direct%26mode=render%26action=download%26mode=render.

4.1 Dictator Experiments Hypotheses

4.1.1 Effect of Taking Option

We hypothesize that giving close to zero is more appropriate in the Take-2 DG and Take-10 DG than in the Standard DG. Formally we expect to reject the following null hypotheses in favor of the alternatives, for $0 \leq X \leq 2$, both for social and personal appropriateness:

$$\begin{aligned} H_0 &: R_{Give\ X}(Take2DG) = R_{Give\ X}(StandardDG) \\ H_1 &: R_{Give\ X}(Take2DG) > R_{Give\ X}(StandardDG), \end{aligned}$$

$$\begin{aligned} H_0 &: R_{Give\ X}(Take10DG) = R_{Give\ X}(StandardDG) \\ H_1 &: R_{Give\ X}(Take10DG) > R_{Give\ X}(StandardDG). \end{aligned}$$

For the Take-2 DG and social appropriateness, this has been verified by Krupka and Weber (2013).

4.1.2 Effect of Surplus Generation

We hypothesize that giving close to zero is more appropriate in the DG with Quiz and the DG with Production than in the Standard DG. Formally we expect to reject the following null hypotheses in favor of the alternatives, for giving $0 \leq X \leq 2$, both for social and personal appropriateness:

$$\begin{aligned} H_0 &: R_{Give\ X}(DGQuiz) = R_{Give\ X}(StandardDG) \\ H_1 &: R_{Give\ X}(DGQuiz) > R_{Give\ X}(StandardDG), \end{aligned}$$

$$
\begin{aligned}
H_0 &: R_{Give\ X}\left(DGProduction\right) = R_{Give\ X}\left(StandardDG\right) \\
H_1 &: R_{Give\ X}\left(DGProduction\right) > R_{Give\ X}\left(StandardDG\right).
\end{aligned}
$$

We are not aware of any test of these hypotheses in the literature.

4.1.3 Effect of Exit Option

We hypothesize that taking the exit option (giving 9 for the dictator and 0 for the recipient) is rated as more appropriate than staying in the game and giving 1 to recipient (leaving 9 for the dictator). We expect to reject the following null hypothesis in favor of the alternative,

$$
\begin{aligned}
H_0 &: R_{Exit}\left(DGExit\right) = R_{Stay,Give\ 1}\left(DGExit\right) \\
H_1 &: R_{Exit}\left(DGExit\right) > R_{Stay,Give\ 1}\left(DGExit\right).
\end{aligned}
$$

For an exit option giving 10 to the dictator and 0 to the recipient, and social appropriateness, this has been verified by Krupka and Weber (2013).

4.1.4 Effect of Information Avoidance

We hypothesise that choosing to become informed, learning that one is in the unaligned interest state, and choosing Y (yielding $5, 5$), is more appropriate than choosing to stay uninformed. Moreover, we hypothesize that choosing to become informed, learning that one is in the unaligned interest state, and choosing X (yielding $6, 1$), is more inappropriate than choosing to stay uninformed. Formally we expect to reject the following null hypotheses in favor of the alternatives,

$$
\begin{aligned}
H_0 &: R_{Informed,\,Y\ in\ unaligned}\left(DGInfo\right) = R_{Uniformed,\,Y}\left(DGInfo\right) \\
H_1 &: R_{Informed,\,Y\ in\ unaligned}\left(DGInfo\right) > R_{Uniformed,\,Y}\left(DGInfo\right),
\end{aligned}
$$

$$
\begin{aligned}
H_0 &: R_{Informed,\,X\ in\ unaligned}\left(DGInfo\right) = R_{Uniformed,\,X}\left(DGInfo\right) \\
H_1 &: R_{Informed,\,X\ in\ unaligned}\left(DGInfo\right) < R_{Uniformed,\,X}\left(DGInfo\right).
\end{aligned}
$$

Krupka & Weber (2008), the working paper version of Krupka and Weber (2013), report results in line with these hypotheses, for the case of social appropriateness.

4.2 Lost Wallet Situation Hypotheses

4.2.1 Inappropriateness of Not Returning

We hypothesize that more people consider keeping the wallet inappropriate than appropriate, regardless of content. Let $f\,(x|m)$ be the frequency of response x when amount of money in the wallet is $m \in \{No, Small, Big\}$. Formally, with the 5-item scale, we expect to reject the following null in favor of the alternative:

$$\begin{aligned} H_0 &: f_{Keep}\,(VI|m) + f_{Keep}\,(SI|m) = f_{Keep}\,(SA|m) + f_{Keep}\,(VA|m) \\ H_1 &: f_{Keep}\,(VI|m) + f_{Keep}\,(SI|m) > f_{Keep}\,(SA|m) + f_{Keep}\,(VA|m)\,. \end{aligned}$$

The hypotheses for the 7-item scale are analogous.

4.2.2 Comparing Inappropriateness of Not Returning when Wallet Contains Small Money or No Money

We hypothesize that not returning the wallet is more inappropriate when the wallet contains a small amount of money than when it contains no money. We can formalize this hypothesis in two ways, either comparing average ratings or comparing the distributions directly. In a direct comparison we compare the frequency of VI-responses. Formally, we expect to reject the following null in favor of the alternative:

$$\begin{aligned} H_0 &: f_{Keep}\,(VI|Small) = f_{Keep}\,(VI|No) \\ H_1 &: f_{Keep}\,(VI|Small) > f_{Keep}\,(VI|No)\,. \end{aligned}$$

When comparing averages we expect to reject the following null in favour of the alternative:

$$\begin{aligned} H_0 &: R_{Keep}\,(Small) = R_{Keep}\,(No) \\ H_1 &: R_{Keep}\,(Small) > R_{Keep}\,(No)\,. \end{aligned}$$

4.2.3 Inappropriateness of not Returning Increasing in Monetary Content

We hypothesize that not returning the wallet is more inappropriate the more money the wallet contains. Formally, we expect to reject the following two null hypotheses in favor of the alternatives:

$$\begin{aligned} H_0 &: f_{Keep}\,(VI|Big) = f_{Keep}\,(VI|Small) \\ H_1 &: f_{Keep}\,(VI|Big) > f_{Keep}\,(VI|Small)\,. \end{aligned}$$

When comparing averages we expect to reject the following null in favour of the alternative:

$$\begin{aligned} H_0 &: R_{Keep}(Big) = R_{Keep}(Small) \\ H_1 &: R_{Keep}(Big) > R_{Keep}(Small). \end{aligned}$$

4.3 Day-Care Center Situation Hypotheses

4.3.1 Inappropriateness of Picking up Late

We hypothesize that more people consider picking up late to be inappropriate than appropriate, regardless of fine. Let $f(x|m)$ be the frequency of response x when the fine is $m \in \{No, Small, Big\}$. Formally, with the 5-item scale, we expect to reject the following null in favor of the alternative:

$$\begin{aligned} H_0 &: f_{Late}(VI|m) + f_{Late}(SI|m) = f_{Late}(SA|m) + f_{Late}(VA|m) \\ H_1 &: f_{Late}(VI|m) + f_{Late}(SI|m) > f_{Late}(SA|m) + f_{Late}(VA|m). \end{aligned}$$

The hypotheses for the 7-item scale are analogous.

4.3.2 Comparing Inappropriateness under Fine and No Fine

We hypothesize that being late is more inappropriate when there is no fine than when there is a small fine. Furthermore, being late is more inappropriate when there is no fine than when there is a big fine. Formally, when comparing distributions directly the following two null hypotheses should be rejected in favor of the alternatives:

$$\begin{aligned} H_0 &: f_{Late}(VI|Small) = f_{Late}(VI|No) \\ H_1 &: f_{Late}(VI|Small) < f_{Late}(VI|No), \end{aligned}$$

$$\begin{aligned} H_0 &: f_{Late}(VI|Big) = f_{Late}(VI|No) \\ H_1 &: f_{Late}(VI|Big) < f_{Late}(VI|No). \end{aligned}$$

When comparing averages we expect to reject the following two null hypotheses should be rejected in favor of the alternatives:

$$\begin{aligned} H_0 &: R_{Late}(Small) = R_{Late}(No) \\ H_1 &: R_{Late}(Small) > R_{Late}(No), \end{aligned}$$

$$
\begin{aligned}
H_0 &: R_{Late}(Big) = R_{Late}(No) \\
H_1 &: R_{Late}(Big) > R_{Late}(No).
\end{aligned}
$$

4.3.3 Comparing Inappropriateness Under Small and Large Fine

We hypothesize that being late is more inappropriate when there is a small fine than when there is a big fine.Formally, we expect to reject the following null in favor of the alternative:

$$
\begin{aligned}
H_0 &: f_{Late}(VI|Big) = f_{Late}(VI|Small) \\
H_1 &: f_{Late}(VI|Big) < f_{Late}(VI|Small).
\end{aligned}
$$

When comparing averages we expect to reject the following null in favor of the alternative:

$$
\begin{aligned}
H_0 &: R_{Late}(Big) = R_{Late}(Small) \\
H_1 &: R_{Late}(Big) > R_{Late}(Small).
\end{aligned}
$$

5 Main Hypotheses Tests

5.1 Dictator Experiments

Table A1. DE stnd. vs DE take: social and personal appropriateness (t-tests)

	Obs (a)	Obs (b)	Mean (a)	Mean (b)	Diff. (a) - (b)	P-value (H0: Diff. = 0, H1: Diff. < 0)
Social appropriateness						
St. DE (0) - DE take 2 (give 0)	148	157	-.649	.01	-.658	0
St. DE (0) - DE take 10 (give 0)	148	135	-.649	.241	-.889	0
St. DE (1) - DE take 2 (give 1)	148	157	-.476	.073	-.55	0
St. DE (1) - DE take 10 (give 1)	148	135	-.476	.241	-.717	0
St. DE (2) - DE take 2 (give 2)	148	157	-.355	.156	-.511	0
St. DE (2) - DE take 10 (give 2)	148	135	-.355	.289	-.644	0
Personal appropriateness						
St. DE (0) - DE take 2 (give 0)	152	125	-.599	-.02	-.579	0
St. DE (0) - DE take 10 (give 0)	152	171	-.599	.14	-.739	0
St. DE (1) - DE take 2 (give 1)	152	125	-.487	.04	-.527	0
St. DE (1) - DE take 10 (give 1)	152	171	-.487	.164	-.651	0
St. DE (2) - DE take 2 (give 2)	152	125	-.391	.12	-.511	0
St. DE (2) - DE take 10 (give 2)	152	170	-.391	.153	-.544	0

Notes. P-value equals to 0 in the last column means that $p < 0.001$.

Table A2. DE stnd. vs DE quiz: social and personal appropriateness (t-tests)

	Obs (a)	Obs (b)	Mean (a)	Mean (b)	Diff. (a) - (b)	P-value (H0: Diff. = 0, H1: Diff. < 0)
Social appropriateness						
St. DE (0) - DE Quiz (0)	148	155	-.649	-.348	-.3	0
St. DE (1) - DE Quiz (1)	148	155	-.476	-.165	-.312	0
St. DE (2) - DE Quiz (2)	148	155	-.355	-.058	-.297	0
Personal appropriateness						
St. DE (0) - DE Quiz (0)	152	150	-.599	-.213	-.385	0
St. DE (1) - DE Quiz (1)	152	150	-.487	-.12	-.367	0
St. DE (2) - DE Quiz (2)	152	150	-.391	-.07	-.321	0

Notes. P-value equals to 0 in the last column means that $p < 0.001$.

Table A3. DE stand. vs DE production: social and personal appropriateness (t-tests)

	Obs (a)	Obs (b)	Mean (a)	Mean (b)	Diff. (a) - (b)	P-value (H0: Diff. = 0, H1: Diff. < 0)
Social appropriateness						
St. DE (0) - DE Prod (0)	148	148	-.649	-.209	-.439	0
St. DE (1) - DE Prod (1)	148	148	-.476	-.084	-.392	0
St. DE (2) - DE Prod (2)	148	148	-.355	-.041	-.314	0
Personal appropriateness						
St. DE (0) - DE Prod (0)	152	152	-.599	.043	-.641	0
St. DE (1) - DE Prod (1)	152	152	-.487	.013	-.5	0
St. DE (2) - DE Prod (2)	152	152	-.391	.053	-.444	0

Notes. P-value equals to 0 in the last column means that $p < 0.001$.

Table A4. DE quiz vs DE production: social and personal appropriateness (t-tests)

	Obs (a)	Obs (b)	Mean (a)	Mean (b)	Diff. (a) - (b)	P-value (H0: Diff. = 0, H1: Diff. < 0)
Social appropriateness						
DE Quiz (0) - DE Prod (0)	155	148	-.348	-.209	-.139	.048
DE Quiz (1) - DE Prod (1)	155	148	-.165	-.084	-.08	.153
DE Quiz (2) - DE Prod (2)	155	148	-.058	-.041	-.018	.405
Personal appropriateness						
DE Quiz (0) - DE Prod (0)	150	152	-.213	.043	-.256	.002
DE Quiz (1) - DE Prod (1)	150	152	-.12	.013	-.133	.042
DE Quiz (2) - DE Prod (2)	150	152	-.07	.053	-.123	.047

Table A5. DE stnd. vs DE exit: social and personal appropriateness (t-tests)

	Obs (a)	Obs (b)	Mean (a)	Mean (b)	Diff. (a) - (b)	P-value (H0: Diff. = 0, H1: Diff. < 0)
Social appropriateness						
DE ex. (give 0) - DE ex. (exit)	161	161	-.73	-.149	-.581	0
DE ex. (give 1) - DE ex. (exit)	161	161	-.475	-.149	-.326	0
Personal appropriateness						
DE ex. (give 0) - DE ex. (exit)	152	152	-.625	-.125	-.5	0
DE ex. (give 1) - DE ex. (exit)	152	152	-.319	-.125	-.194	.004

Notes. P-value equals to 0 in the last column means that $p < 0.001$.

Table A6. DWK treatment: social and personal appropriateness (t-tests)

	Obs (a)	Obs (b)	Mean (a)	Mean (b)	Diff. (a) - (b)	P-value (H0: Diff. = 0, H1: Diff. < 0)
Social appropriateness						
Uninf. - A chooses Y, inf.	150	150	.135	.823	-.688	0
A chooses X, inf. - Uninf.	150	150	-.513	.135	-.648	0
Personal appropriateness						
Uninf. - A chooses Y, inf.	150	150	.267	.787	-.52	0
A chooses X, inf. - Uninf.	150	150	-.493	.267	-.76	0

Notes. P-value equals to 0 in the last column means that $p < 0.001$.

5.2 Day-Care Center Situation Hypotheses

5.2.1 Inappropriateness of Picking up Late

Table A7. Inppropriateness of arriving late (t-tests)

	Obs	Proportion	Std. Error	P-value (H0: Pr=0.5, H1: Pr>0.5)
Social appropriateness	200	.93	.018	0
Personal appropriateness	194	.9175	.0198	0

Notes. P-value equals to 0 in the last column means that p < 0.001.

5.2.2 Comparing Inappropriateness Under Fine and No Fine

Table A8. Comparing inappropriateness of arriving late under (small or big) fine or no fine (t-tests)

	Mean (1)	Mean (2)	Diff.	P-valule (H0: diff=0, H1: diff<0)
Social appropriateness				
No fine (15) - Small fine (15)	-.2766	-.0736	-.203	0
No fine (30) - Small fine (30)	-.7183	-.5508	-.1675	0
No fine (15) - Big fine (15)	-.2766	-.0635	-.2132	0
No fine (30) - Big fine (30)	-.7183	-.4365	-.2817	0
Personal appropriateness				
No fine (15) - Small fine (15)	-.3686	-.2036	-.1649	0
No fine (30) - Small fine (30)	-.732	-.5928	-.1392	0
No fine (15) - Big fine (15)	-.3686	-.1495	-.2191	0
No fine (30) - Big fine (30)	-.732	-.4845	-.2474	0

Notes. P-value equals to 0 in the last column means that p < 0.001.

5.2.3 Comparing Inappropriateness Under Small and Large Fine

Table A9. Comparing inappropriateness of arriving late under small and large fine - 15 mins of delay (t-tests)

	Mean (1)	Mean (2)	Diff.	P-valule (H0: diff=0, H1: diff<0)
Social appropriateness				
Small fine (15) - Big fine (15)	-.0736	-.0635	-.0102	.3778
Small fine (30) - Big fine (30)	-.5508	-.4365	-.1142	.0003
Personal appropriateness				
Small fine (15) - Big fine (15)	-.2036	-.1495	-.0541	.0563
Small fine (30) - Big fine (30)	-.5928	-.4845	-.1082	.0006

5.3 Lost Wallet Situations

5.3.1 Inappropriateness of Not Returning

Table A10. Inappropriateness of not returning (5 scale, t-tests)

	Obs	Proportion	Std. Error	P-value (H0: Pr=0.5, H1: Pr>0.5)
Social appropriateness	200	.975	.011	0
Personal appropriateness	194	.9948	.0051	0

Notes. P-value equals to 0 in the last column means that p < 0.001.

Table A11. Inappropriateness of not returning (7 scale, t-tests)

	Obs	Proportion	Std. Error	P-value (H0: Pr=0.5, H1: Pr>0.5)
Social appropriateness	201	1	0	0
Personal appropriateness	200	.995	.005	0

Notes. P-value equals to 0 in the last column means that p < 0.001.

5.3.2 Comparing Inappropriateness of Not Returning when Wallet Contains Small Money or No Money

Table A12. Comparing the inappropiateness of not returning when wallet contains small money or no money (5 scale, t-tests)

	Mean (1)	Mean (2)	Diff.	P-value (H0: diff=0, H1:diff<0)
Social appropriateness				
Small money - No money	-.9167	-.8965	-.0202	.0789
Personal appropriateness				
Small Money - No money	-.9433	-.9046	-.0387	.008

Table A13. Comparing inappropriateness of not returning the wallet when the wallet contains small money or no money (7 scale, t-tests)

	Mean (1)	Mean (2)	Diff.	P-value (H0: diff=0, H1:diff<0)
Social appropriateness				
Small money - No money	-.8773	-.8275	-.0498	.0129
Personal appropriateness				
Small Money - No money	-.8833	-.8633	-.02	.145

5.3.3 Inappropriateness of Not Returning Increasing in Monetary Content

Table A14. Inappropiateness of not returning increasing in monetary content (5 scale, t-tests)

	Mean (1)	Mean (2)	Diff.	P-value (H0: diff=0, H1:diff<0)
Social appropriateness				
Big money - Small money	-.9369	-.9167	-.0202	.0853
Personal appropriateness				
Big money - Small Money	-.9253	-.9433	.018	.8623

Table A15. Inappropriateness of not returning increasing in monetary content (7 scale, t-tests)

	Mean (1)	Mean (2)	Diff.	P-value (H0: diff=0, H1:diff<0)
Social appropriateness				
Big money - Small money	-.9171	-.8773	-.0398	.0086
Personal appropriateness				
Big money - Small Money	-.9033	-.8833	-.02	.0956

6 Order Effects

We test order effects in two ways, at the level of individual questions (one for each action), and at the level of scenarios (a dictator game, a lost wallet situation, or a day-care center situation).

We test order effects at the level of question (actions) as follows. For each question (i.e. each action) in each scenario, we compare (i) average rating by subjects who saw that scenario as the first scenario in the set of three scenarios to which it belong, with (ii) average rating by all other subjects who saw that scenario (under the same number of rating alternatives and the same mode of appropriateness). We use a t-test to compare the average ratings of these two groups.

We test order effects at the level of scenarios as follows. For each action in a given scenario, we code the individual responses numerically as we do to calculate averages. Then we use the resulting set of numerical variables (one for each question in the given scenario) as an explanatory variable, in a regression to predict whether the subject in question saw the scenario first or not. We use an F-test to determine whether the coefficients are jointly significant.

6.1 Order Effects in Dictator Experiments

With both approaches we find clear evidence of order effect in the Take-2 Dictator experiment, for social as well as personal appropriateness. Subjects who face this game first find it more appropriate to give 0-4, than those who face this game later. In the Take-10 Dictator experiment there is a statistically significant order effect only for the option to give 0. Subjects who face this game first find it more appropriate to give 0, than those who face this game later. We also find some evidence of order effects in the Dictator experiment with Exit option and the Dictator experiment with Production. Since the results of both approaches cohere, we only reports the more detailed results from the first approach, and only for the Take-2 Dictator experiment, where there is a more pronounced effect. In addition we provide graphs of average rating for all dictator experiments using only data from subjects who saw the game first.

Table A16. DE take 2: order effects in social appropriateness (t-tests)

	Obs.	Mean (others)	Mean (seen first)	Diff.	P-value
DE take 2 (-2)	157	-0.5000	-0.5833	0.0833	0.7765
DE take 2 (-1)	157	-0.6783	-0.5357	-0.1425	0.0462
DE take 2 (0)	157	-0.2826	0.0595	-0.3421**	0.0017
DE take 2 (1)	157	-0.1348	0.4048	-0.5395***	0.0000
DE take 2 (2)	157	-0.0565	0.4286	-0.4851***	0.0000
DE take 2 (3)	157	0.0478	0.4524	-0.4046***	0.0002
DE take 2 (4)	157	0.3087	0.5119	-0.2032*	0.0160
DE take 2 (5)	157	0.8000	0.7381	0.0619	0.8078
DE take 2 (6)	157	0.4391	0.4405	-0.0013	0.4944
DE take 2 (7)	157	0.3174	0.3690	-0.0517	0.3174
DE take 2 (8)	157	0.2478	0.3095	-0.0617	0.3006
DE take 2 (9)	157	0.1609	0.2500	-0.0891	0.2513
DE take 2 (10)	157	0.1478	0.2143	-0.0665	0.3201

* p<0.05, ** p<0.01, *** p<0.001

Table A17. DE take 2: order effects in personal appropriateness (t-tests)

	Obs.	Mean (others)	Mean (seen first)	Diff.	P-value
DE take 2 (-2)	125	-0.4890	-0.5294	0.0404	0.6401
DE take 2 (-1)	125	-0.5989	-0.4118	-0.1871	0.0381
DE take 2 (0)	125	-0.0879	0.1618	-0.2497	0.0379
DE take 2 (1)	125	-0.0604	0.3088	-0.3693**	0.0019
DE take 2 (2)	125	0.0330	0.3529	-0.3200**	0.0036
DE take 2 (3)	125	0.0934	0.2794	-0.1860	0.0548
DE take 2 (4)	125	0.3132	0.3529	-0.0398	0.3507
DE take 2 (5)	125	0.6868	0.5294	0.1574	0.9475
DE take 2 (6)	125	0.3022	0.0735	0.2287*	0.9772
DE take 2 (7)	125	0.0549	0.0294	0.0255	0.5811
DE take 2 (8)	125	0.0000	-0.0735	0.0735	0.7173
DE take 2 (9)	125	-0.0604	-0.1471	0.0866	0.7357
DE take 2 (10)	125	-0.1648	-0.1765	0.0116	0.5303

* p<0.05, ** p<0.01, *** p<0.001

6.2 Dictator experiment ratings accounting for the order of games

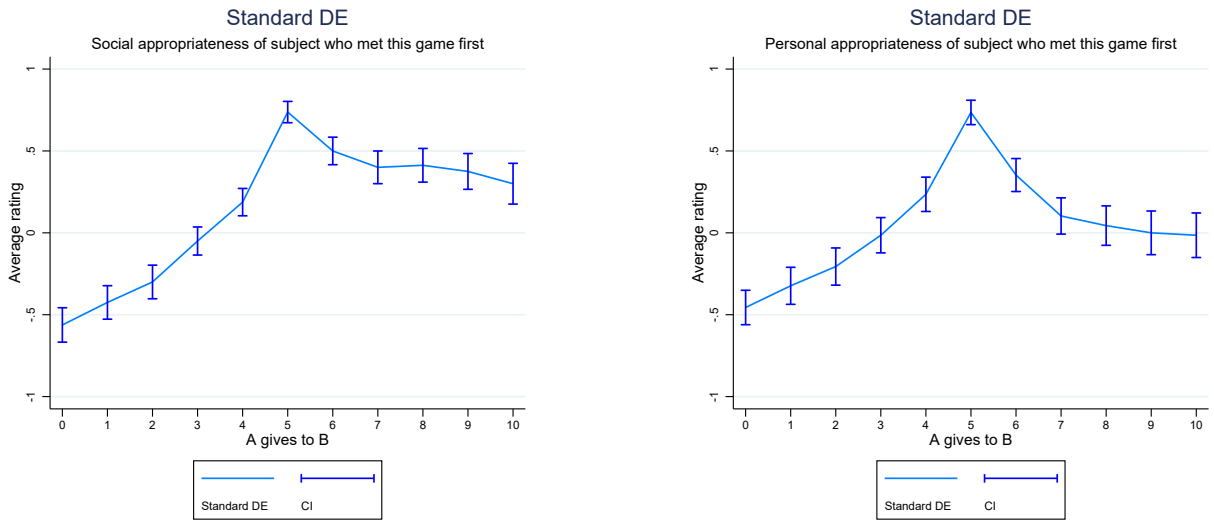


Figure A1. Average social (right) and personal (left) appropriateness ratings in the standard DE, for subject who met this game first (95% confidence intervals)

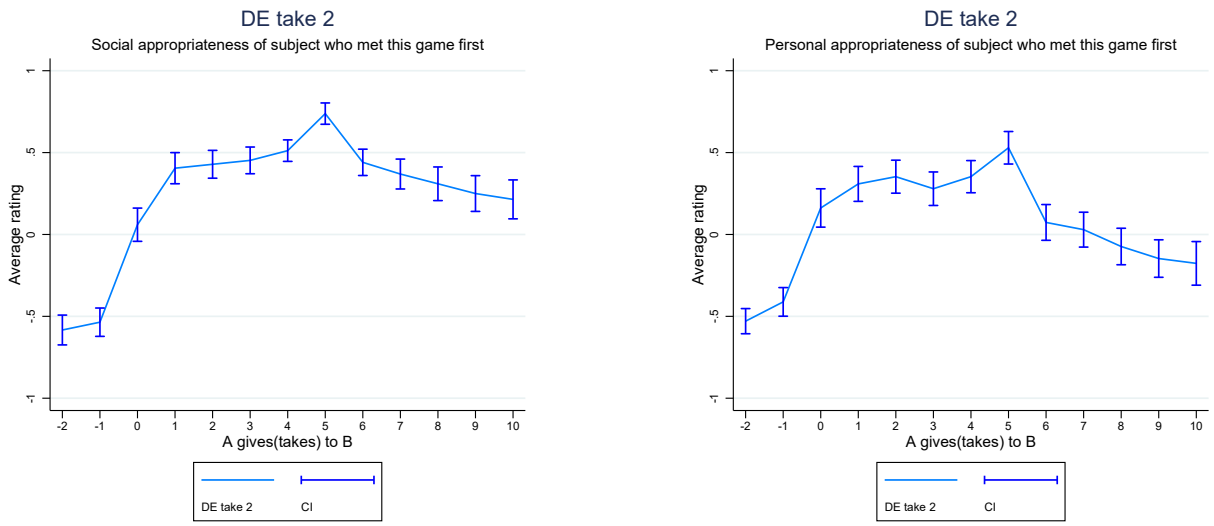
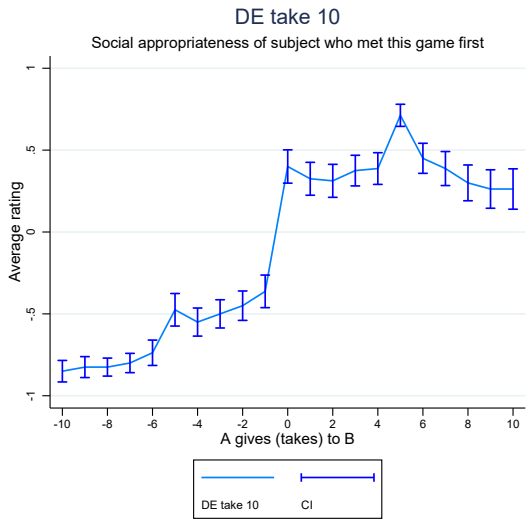


Figure A2. Average social (right) and personal (left) appropriateness ratings in the Take 2 DE, for subject who met this game first (95% confidence intervals)



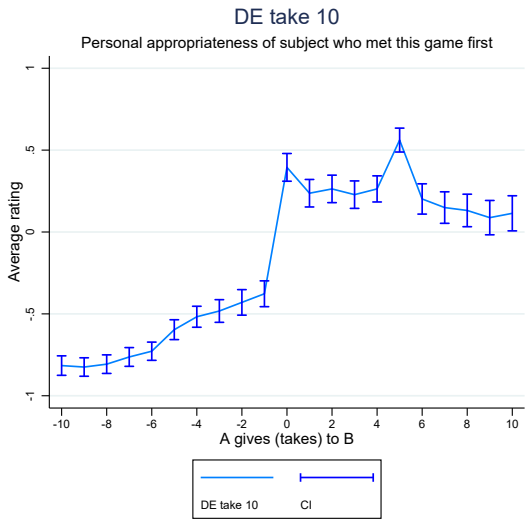
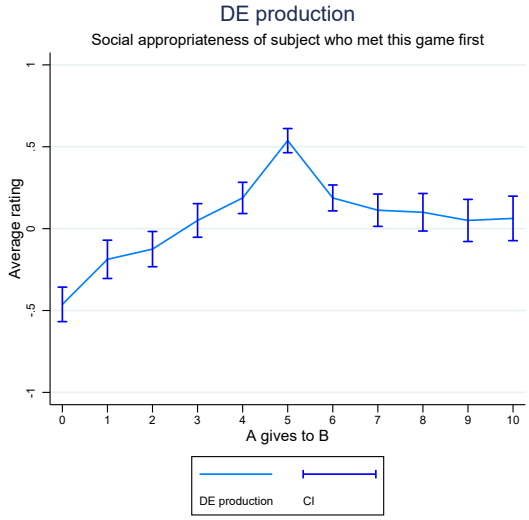


Figure A3. Average social (right) and personal (left) appropriateness ratings in the Take 10 DE, for subject who met this game first (95% confidence intervals)



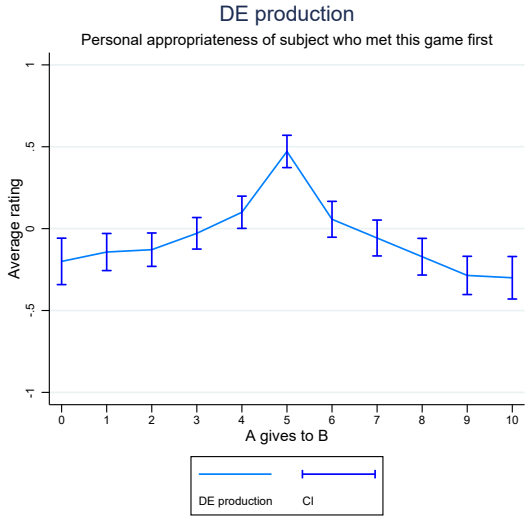


Figure A4. Average social (right) and personal (left) appropriateness ratings in the DE with production, for subject who met this game first (95% confidence intervals)

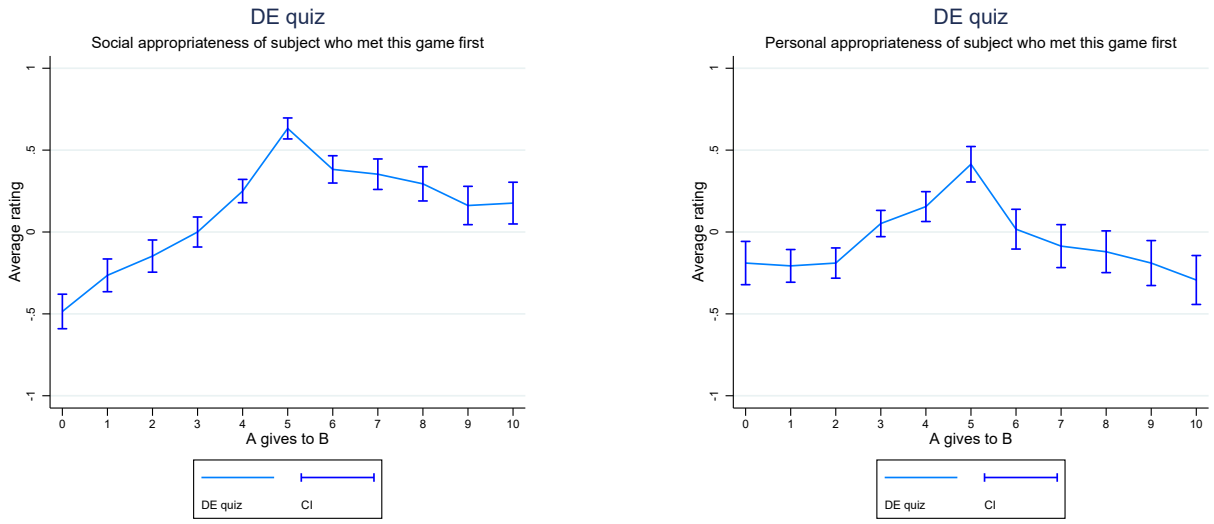


Figure A5. Average social (right) and personal (left) appropriateness ratings in the DE with quiz, for subject who met this game first (95% confidence intervals)

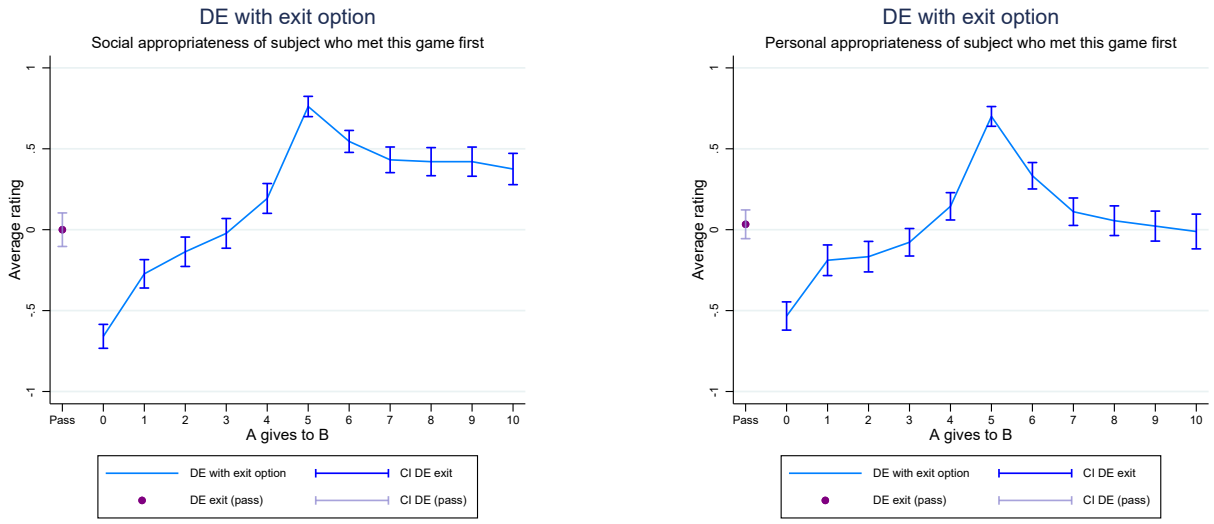


Figure A6. Average social (right) and personal (left) appropriateness ratings in the DE with exit option, for subject who met this game first (95% confidence intervals)

6.3 Order Effects in Lost wallet Situations

Since there are two questions per scenario (version of the lost wallet situation) we only test for order effects at the level of questions / actions. We do not find any evidence of order effects. Results are available upon request.

7 Instructions for Experiments

Here we reproduce a sample of screenshots from the instructions for the experiments eliciting social and personal appropriateness ratings in the dictator experiments. Complete instructions for all elicitation tasks (all dictator experiments, the lost wallet situations, and the day-care center situations), for both personal and social appropriateness are available at https://sites.google.com/site/karlerikmohlin/home.

Note that we use a lost wallet situation, taken from Krupka and Weber (2013), as an example in these experimental instructions. For our experiments eliciting appropriateness in the lost wallet and day-care situations we use a different introductory example, involving donations to a beggar.

7.1 Social Appropriateness in Dictator Experiments: Introduction

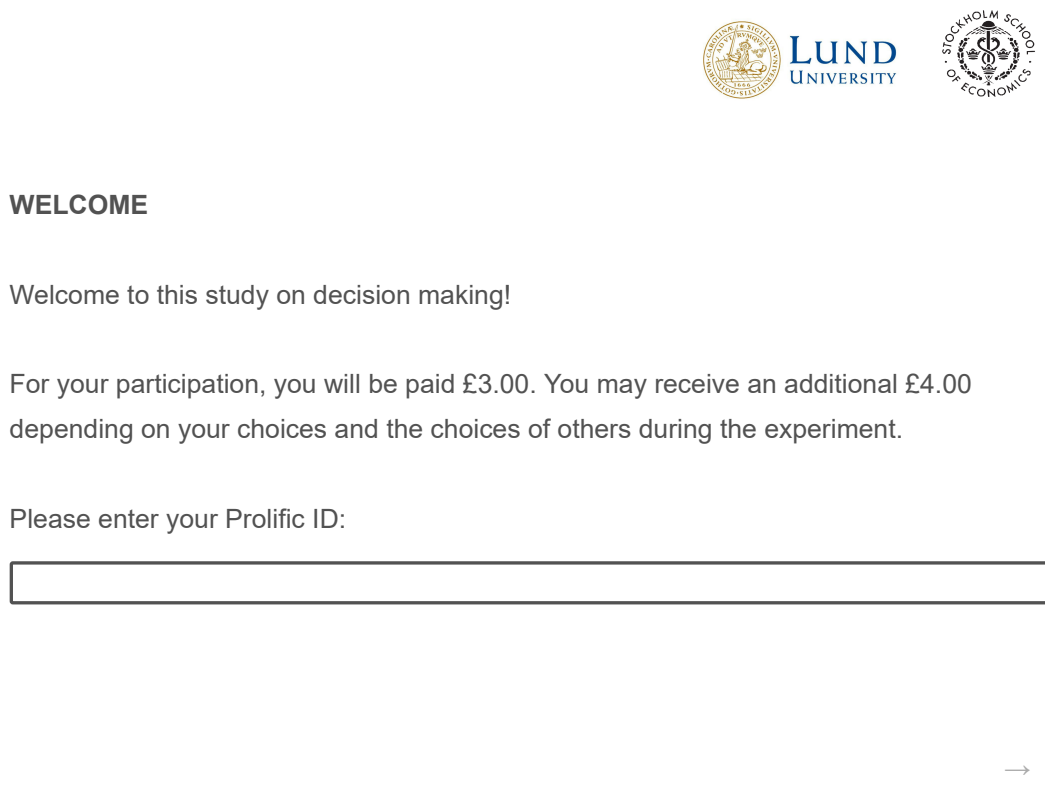


Figure A7. Introduction to Social Appropriateness in Dictator Experiments: Screen 1

YOUR TASK

In the following you will read descriptions of different situations. In each situation there is one person who must make a choice between different actions. After you have read the description of each situation, we ask you to evaluate the different actions.

For each of the possible actions, we ask you to what degree you think that taking that action would be **socially appropriate** or **socially inappropriate**.

By socially **appropriate**, **we mean** behavior that most people agree is the "**correct**" or "**proper**" or "**ethical**" thing to do. Another way to think about what we mean is that if a person were to select a socially **inappropriate** choice, then someone else might be **angry** at the person for doing so.

We will now go through an example.

→

Figure A8. Introduction to Social Appropriateness in Dictator Experiments: Screen 2

EXAMPLE

Individual A is at a local coffee shop. While there, Individual A notices that someone has left a wallet at one of the tables. Individual A has four possible choices: take the wallet, ask others nearby if the wallet belongs to them, leave the wallet where it is, or give the wallet to the shop manager.

The table below presents the possible choices available to Individual A. Please rate the social appropriateness of each of the choices.

Recall that by socially appropriate, we mean behavior that most people agree is the "correct" or "proper" or "ethical" thing to do. Another way to think about what we mean is that if a person were to select a socially inappropriate choice, then someone else might be angry at the person for doing so.

Please rate the social appropriateness of Individual A's possible choices

	Very socially inappropriate	Somewhat socially inappropriate	Neither socially appropriate nor inappropriate	Somewhat socially appropriate	Very socially appropriate
Take the wallet	○	○	○	○	○
Ask others nearby if the wallet belongs to them	○	○	○	○	○
Leave the wallet where it is	○	○	○	○	○
Give the wallet to the shop manager	○	○	○	○	○

→

Figure A9. Introduction to Social Appropriateness in Dictator Experiments: Screen 3

THE SITUATIONS THAT FOLLOW

On the following pages, there are a number of situations described. **All situations deal with decisions that "Individual A" might have to make.**

For each situation, we ask you to indicate to what degree each possible choice available to Individual A is socially appropriate or socially inappropriate.

→

Figure A10. Introduction to Social Appropriateness in Dictator Experiments: Screen 4

ADDITIONAL PAYMENT

At the end of the experiment today, we will select one of the situations. For the selected situation, we will also randomly select one of the possible choices that Individual A could make. Thus, we will select both a situation and one possible choice at random.

For the choice selected, we will determine which response was selected by most people here today. **If you give the same response as that given by most other people, then you will receive an additional £4.00**.

For instance, suppose we were to select the example situation above and the possible choice “Leave the wallet where it is”. Moreover, suppose that most other people taking this survey had selected the response “somewhat socially inappropriate” to this question. If your response had also been “somewhat socially inappropriate,” then you would receive £4.00, in addition to the £3.00 participation reward. Otherwise, you would receive only the £3.00 participation reward.

This means that in order to maximize your chances of earning the additional £4.00 you should try to select the response alternative that most other people select.

Note that all participants in this study were born in the US and currently live in the US.

→

Figure A11. Introduction to Social Appropriateness in Dictator Experiments: Screen 5

YOUR CONSENT

The information that you give in the study will be handled confidentially. Your data will be anonymous which means that your name will not be collected or linked to the data.

The data gathered through this study will only be used for the purpose of academic research.

If you agree to participate, please be aware that you are free to withdraw at any point throughout the duration of the experiment. However, in case you withdraw, you receive no payment.

Do you agree to participate in the research study described above?

- ○ I agree to participate
- ○ I do not want to participate

→

Figure A12. Introduction to Social Appropriateness in Dictator Experiments: Screen 6

7.2 Social Appropriateness: Dictator Experiment with Take-2 Option

SITUATION

Suppose that Individual A participates in an experiment on decision-making. Individual A is randomly paired with another Individual in the experiment, Individual B. The pairing is anonymous, meaning that neither individual will ever know the identity of the other individual with whom he or she is paired.

In the experiment, Individual A will make a choice, the experimenter will record this choice, and then both Individual A and Individual B will be informed of the choice and paid money based on the choice made by Individual A, as well as a small payment for participation. Suppose that neither individual will receive any other money for participating in the experiment.

Individual A will receive $10. Individual A will then have the opportunity to give any of this $10 to Individual B, or to take up to $2 from Individual B's payment for participation.

For instance, Individual A may decide to take $1 from Individual B and keep the $10 for him or herself. Or Individual A may decide to give all $10 to Individual B and not take any money. Individual A may also choose to give any other amount between $0 and $10 to Individual B or to take any amount between $0 and $2 from Individual B. This choice will determine how much money each will receive, privately and in cash, at the end of the experiment.

YOUR TASK

The table below presents a list of the possible choices available to Individual A, consisting of different amounts of money that he/she can take from Individual B or give to Individual B. Please rate the social appropriateness of these choices..

Recall that by socially appropriate, we mean behavior that most people agree is the “correct” or “proper” or “ethical” thing to do. Another way to think about what we mean is that if an Individual were to select a socially inappropriate choice, then someone else might be angry at the Individual for doing so.

Remember: when we select a situation and an action for payment, you will only receive the additional £4 if your response is the same as the most frequent response made by other participants in this survey.

	Very socially inappropriate	Somewhat socially inappropriate	Neither socially appropriate nor inappropriate	Somewhat socially appropriate	Very socially appropriate
Take $2 from Individual B (A gets $12, B gets $8)	○	○	○	○	○
Take $1 from Individual B (A gets $11, B loses $1)	○	○	○	○	○
Neither Give nor Take anything (A gets $10, B gets $0)	○	○	○	○	○
Give $1 to Individual B (A gets $9, B gets $1)	○	○	○	○	○
Give $2 to Individual B (A gets $8, B gets $2)	○	○	○	○	○
Give $3 to Individual B (A gets $7, B gets $3)	○	○	○	○	○
Give $4 to Individual B (A gets $6, B gets $4)	○	○	○	○	○

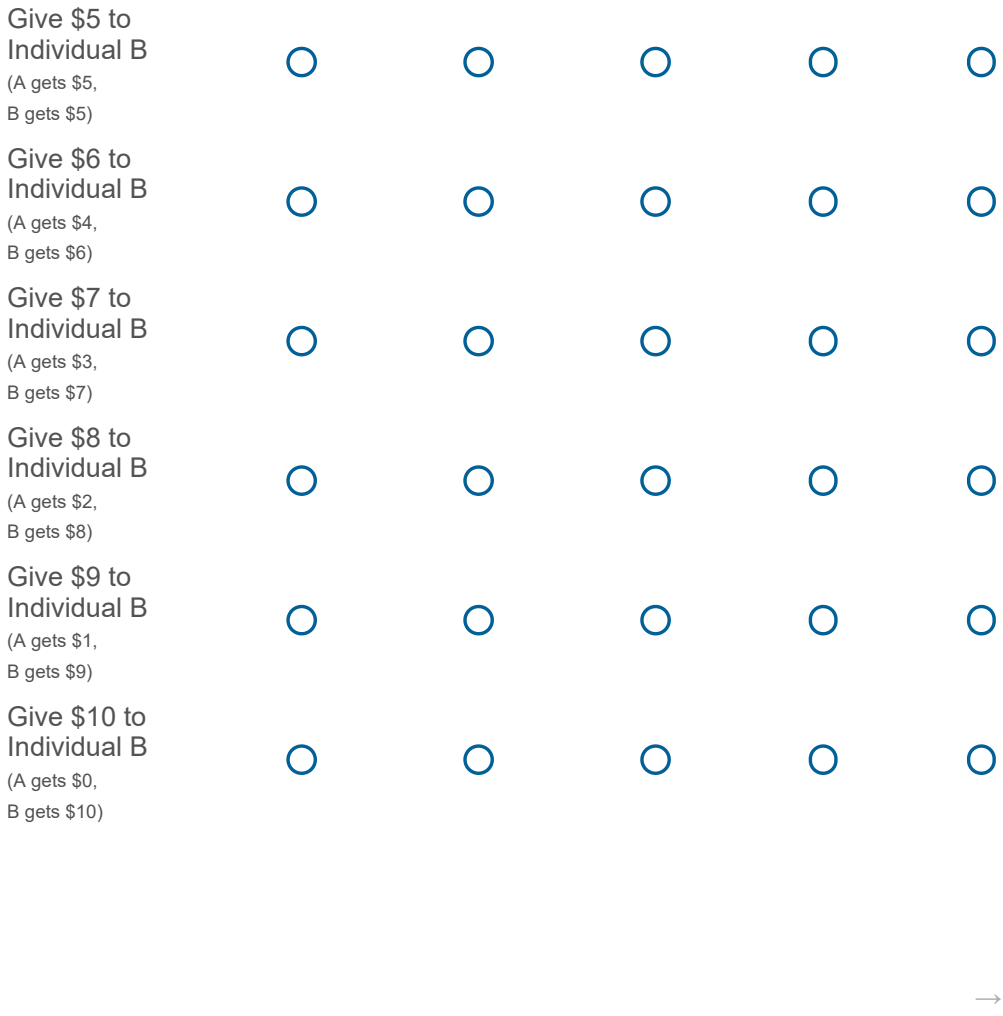


Figure A13. Social Appropriateness: Dictator Experiment with Take-2 Option

7.3 Personal Appropriateness in Dictator Experiments: Introduction

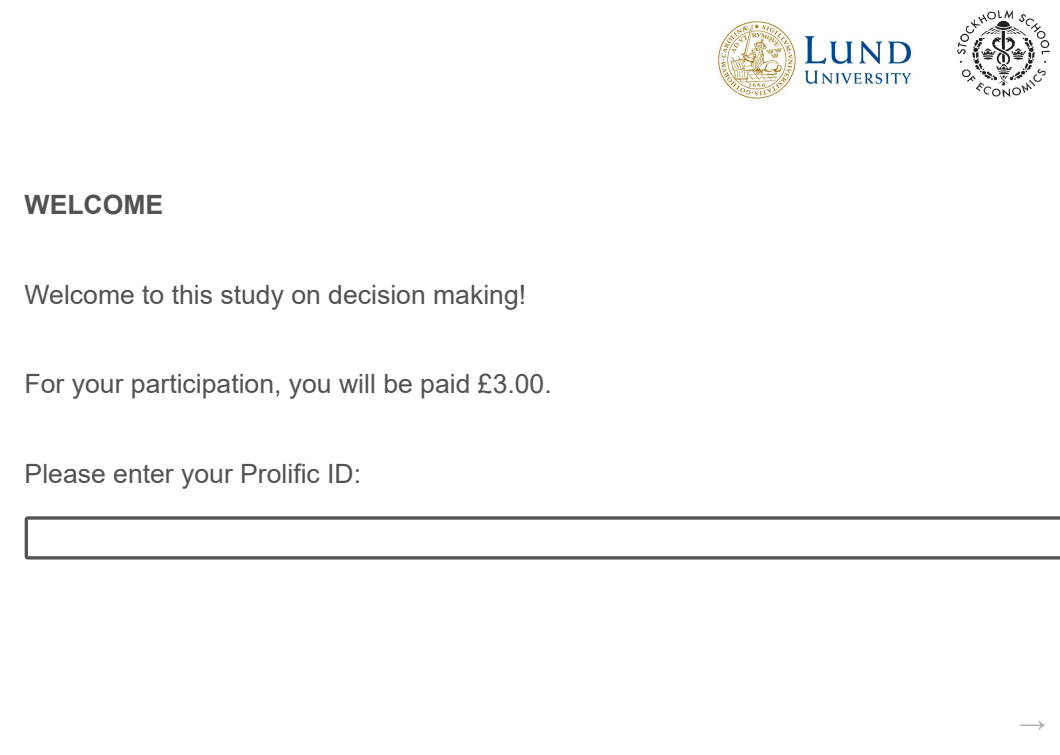


Figure A14. Introduction to Personal Appropriateness in Dictator Experiments: Screen 1

YOUR TASK

In the following you will read descriptions of different situations. In each situation there is one person who must make a choice between different actions. After you have read the description of each situation, we ask you to evaluate the different actions.

For each of the possible actions, we ask you to what degree you think that taking that action would be **appropriate** or **inappropriate**.

By **appropriate, we mean** the behavior that you personally would consider to be the "**correct**" or "**proper**" or "**ethical**" thing to do. Another way to think about what we mean is that if a person were to select an **inappropriate** choice, then you might be **angry** at the person for doing so. We are interested in **your personal opinion, independently of the opinion of others.**

We will now go through an example.

→

Figure A15. Introduction to Personal Appropriateness in Dictator Experiments: Screen 2

EXAMPLE

Individual A is at a local coffee shop. While there, Individual A notices that someone has left a wallet at one of the tables. Individual A has four possible choices: take the wallet, ask others nearby if the wallet belongs to them, leave the wallet where it is, or give the wallet to the shop manager.

The table below presents the possible choices available to Individual A. Please rate the appropriateness of each of the choices.

Recall that by appropriate, we mean the behavior that you personally would consider to be the "correct" or "proper" or "ethical" thing to do. Another way to think about what we mean is that if a person were to select an inappropriate choice, then you might be angry at the person for doing so. We are interested in your personal opinion, independently of the opinion of others.

Please rate the social appropriateness of Individual A's possible choices

	Very inappropriate	Somewhat inappropriate	Neither appropriate nor inappropriate	Somewhat appropriate	Very appropriate
Take the wallet	○	○	○	○	○
Ask others nearby if the wallet belongs to them	○	○	○	○	○
Leave the wallet where it is	○	○	○	○	○
Give the wallet to the shop manager	○	○	○	○	○

→

Figure A16. Introduction to Personal Appropriateness in Dictator Experiments: Screen 3

THE SITUATIONS THAT FOLLOW

On the following pages, there are a number of situations described. **All situations deal with decisions that "Individual A" might have to make.**

For each situation, we ask you to indicate to what degree each possible choice available to Individual A is appropriate or inappropriate.

→

Figure A17. Introduction to Personal Appropriateness in Dictator Experiments: Screen 4

YOUR CONSENT

The information that you give in the study will be handled confidentially. Your data will be anonymous which means that your name will not be collected or linked to the data.

The data gathered through this study will only be used for the purpose of academic research.

If you agree to participate, please be aware that you are free to withdraw at any point throughout the duration of the experiment. However, in case you withdraw, you receive no payment.

Do you agree to participate in the research study described above?

- ○ I agree to participate
- ○ I do not want to participate

→

Figure A18. Introduction to Personal Appropriateness in Dictator Experiments: Screen 5

7.4 Personal Appropriateness: Dictator Game with Take-2 Option

SITUATION

Suppose that Individual A participates in an experiment on decision-making. Individual A is randomly paired with another Individual in the experiment, Individual B. The pairing is anonymous, meaning that neither individual will ever know the identity of the other individual with whom he or she is paired.

In the experiment, Individual A will make a choice, the experimenter will record this choice, and then both Individual A and Individual B will be informed of the choice and paid money based on the choice made by Individual A, as well as a small payment for participation. Suppose that neither individual will receive any other money for participating in the experiment.

Individual A will receive $10. Individual A will then have the opportunity to give any of this $10 to Individual B, or to take up to $2 from Individual B's payment for participation.

For instance, Individual A may decide to take $1 from Individual B and keep the $10 for him or herself. Or Individual A may decide to give all $10 to Individual B and not take any money. Individual A may also choose to give any other amount between $0 and $10 to Individual B or to take any amount between $0 and $2 from Individual B. This choice will determine how much money each will receive, privately and in cash, at the end of the experiment.

YOUR TASK

The table below presents a list of the possible choices available to Individual A, consisting of different amounts of money that he/she can take from Individual B or give to Individual B. Please rate the appropriateness of these choices..

Recall that by appropriate we mean the behavior that you personally would consider to be the "correct" or "proper" or "ethical" thing to do. Another way to think about what we mean is that if a person were to select an inappropriate choice, then you might be angry at the person for doing so. We are interested in your personal opinion, independently of the opinion of others.

	Very inappropriate	Somewhat inappropriate	Neither appropriate nor inappropriate	Somewhat appropriate	Very appropriate
Take $2 from Individual B (A gets $12, B gets $8)	○	○	○	○	○
Take $1 from Individual B (A gets $11, B loses $1)	○	○	○	○	○
Neither Give nor Take anything (A gets $10, B gets $0)	○	○	○	○	○
Give $1 to Individual B (A gets $9, B gets $1)	○	○	○	○	○
Give $2 to Individual B (A gets $8, B gets $2)	○	○	○	○	○
Give $3 to Individual B (A gets $7, B gets $3)	○	○	○	○	○
Give $4 to Individual B (A gets $6, B gets $4)	○	○	○	○	○

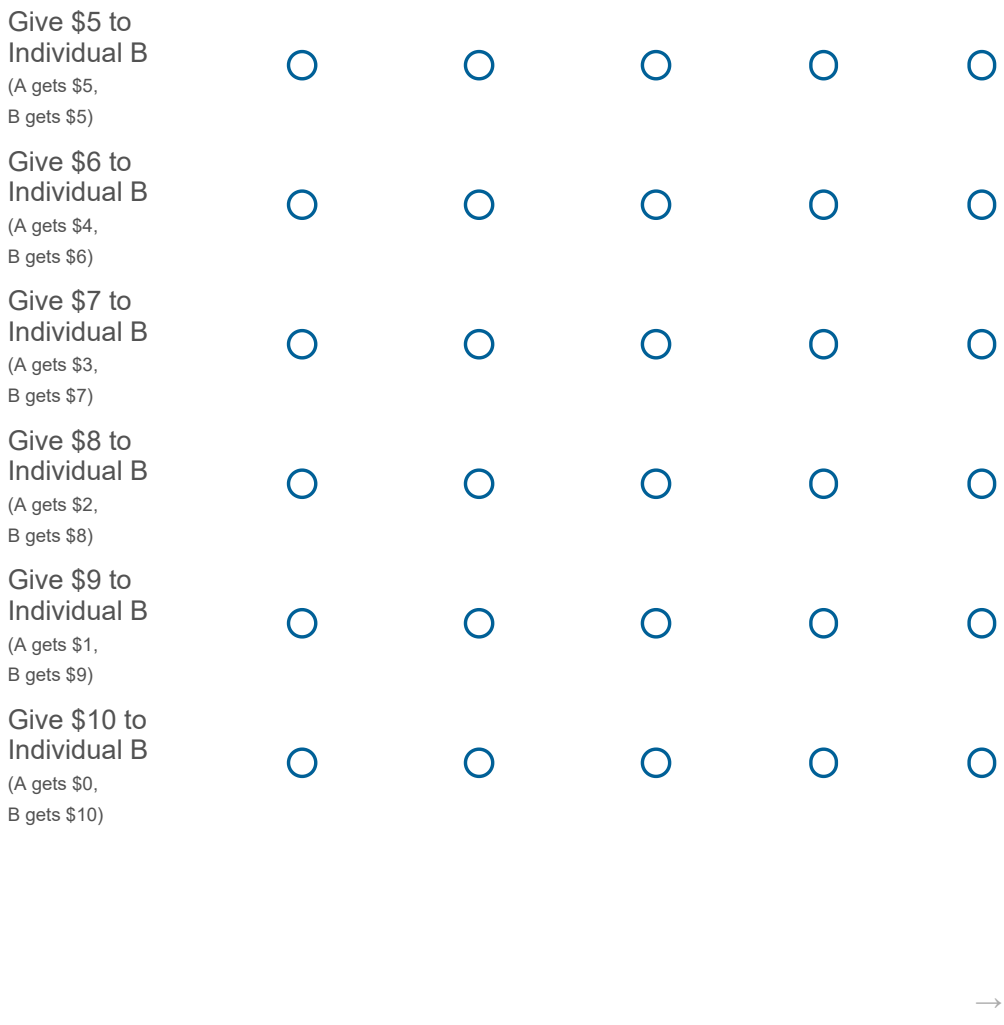


Figure A19. Personal Appropriateness: Dictator Experiment with Take-2 Option