A Theory of Decency*

Tore Ellingsen† Erik Mohlin‡

January 28, 2019

Abstract

We develop a formal model of decency. Shared values and understandings give rise to social norms. While norms may mandate collectively optimal behavior, they need not do so. Furthermore, behavior can be affected by social values even if it stops short of norm compliance. Seeking stronger predictions, we propose a structural model of social values; society endorses efficiency and equality, but condemns ill-gotten gains. The model implies that decent people will tend to avoid situations that encourage pro-social behavior. It also rationalizes the existence of willful ignorance, intention-based negative reciprocity, and betrayal aversion.

JEL Codes: D91, Z13

Keywords: Culture, Norms, Situations, Social Context, Social Preferences

*An earlier draft was entitled Situations and Norms. We thank Sandro Ambuehl, Björn Bartling, Ernst Fehr, Erik Gaard Kristiansen, Zoltán Rácz, Felix Schafmeister, Christian Schultz, Peter Norman Sørensen, Mark Voorneveld, Roberto Weber, and especially Klaus Schmidt, for helpful comments. Ellingsen gratefully acknowledges financial support from the Torsten and Ragnar Söderberg Foundation. Mohlin gratefully acknowledges financial support from the Swedish Research Council (Grant 2015-01751), and the Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellowship 2016-0156).

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

‡Address: Department of Economics, Lund University, Tycho Brahes väg 1, 220 07 Lund, Sweden. E-mail: erik.mohlin@nek.lu.se.

1 Introduction

> *Many men behave very decently, and through the whole of their lives avoid any considerable degree of blame, who yet, perhaps, never felt the sentiment upon the propriety of which we found our approbation of their conduct, but acted merely from a regard to what they saw were the established rules of behaviour.*
>
> Adam Smith
>
> The Theory of Moral Sentiments (1790, Chapter 5, Paragraph 1.)

Why do we give to charity? Why do we tip? Why do we pay taxes that we might easily have avoided? Why do we help colleagues and friends even when we understand that they will be unable to reciprocate? Why do we sometimes incur personal costs in order to punish or harm others? That is, why do we ever pursue social goals instead of our own material well-being?[1]

One reason is *passion*. We are genuinely kind or spiteful, taking joy from others' pleasure or pain. Another reason is *decency*. We feel a duty to act kindly or spitefully.

At first sight, these two reasons may seem similar, but they are not. The altruistic person will cherish opportunities to behave altruistically. By contrast, the decent person may prefer to forgo those opportunities whenever duties are not thereby violated. For example, she might be charitable when faced with a fundraiser, yet take pains to avoid the fundraising drive. Such reluctant charity has recently been documented in field studies by, among others, DellaVigna, List, and Malmendier (2012), Andreoni, Rao, and Trachtman (2017), and Exley and Petrie (2018), building on earlier laboratory studies by Dana, Cain, and Dawes (2006), Broberg, Ellingsen and Johannesson (2007), and Lazear, Malmendier, and Weber, (2012). As noted by Mansbridge (1998), the distinction between passion and duty has been occupying philosophers and social scientists for ages. It is also recognized by personality research. According to standard definitions of "big-5" personality traits, altruism is a facet of Agreeableness whereas dutifulness is a facet of Conscientiousness; see, e.g., McCrae and Costa (2003).

In this paper, our main purpose is to construct a simple and portable formal model of decency and to demonstrate that it offers a unified account for a wide variety of experimental findings that defy standard passion-based models. In addition to reluctant charity, the model explains willful ignorance of externalities (Dana, Weber, and Kuang, 2007; Grossman, 2014; Feiler, 2014), intention-based negative reciprocity (Blount, 1995; Falk, Fehr and Fischbacher,

[1] A possible response to this question is to deny the premise. Maybe we do promote our own material well-being in these cases too. We may be afraid that a selfish act hurts us by causing social contagion (Kandori, 1991). We may even hold "magic beliefs" that if we fail to cooperate, bad consequences will immediately follow (Shafir and Tversky, 1992). While both these effects may matter, the evidence that we survey below indicates that other effects are frequently at play.

2003), and betrayal aversion (Bohnet and Zeckhauser, 2004; Bohnet et al, 2008).

The field experiment of Andreoni, Rao, and Trachtman (2017) illustrates many of our concepts. There, Salvation Army officers are randomly placed outside one or both of the entrances to a supermarket, more or less loudly soliciting charitable donations from shoppers. If shoppers primarily give out of passion, the solicitor's presence at only one door would increase traffic through that door. If shoppers primarily give out of duty, the solicitor's presence would instead decrease traffic through that door and increase traffic through the other door. The study finds that avoidance dominates, with some shoppers taking substantial detours in order to avoid passing by a loud Salvation Army officer. That is, much of the charitable giving seems to be caused by decency rather than passion. In a nutshell, our model rationalizes this finding through its implication that people prefer to be in a situation where they feel less social pressure to act generously. The model also rationalizes the related laboratory finding that people who tend to be more charitable when the situation is inescapable are also more likely to opt out when possible (Lazear, Malmendier, and Weber, 2012).

Beyond explaining the moral behavior of individuals, there are more fundamental reasons why social scientists should distinguish decent behavior from passionate behavior. From a positive perspective, it helps us understand culture. Decency is shaped by powerful cultural forces. Instilling decency is an integral task of many roles and occupations. Parents, teachers, politicians, authors, and managers foist social understandings and values on their children, pupils, voters, readers, and organization members. In comparison with innate moral passions, decency is thus more immediately tied to cultural variation in moral behavior.[2] From a normative perspective, this accentuates the question of how societies can engineer their moral values to obtain other goals, such as material and psychological well-being. As the model makes clear, decency is constraining. Thus, utilitarian welfare calculations associated with such moral engineering should take into account not only the social benefits that decent behavior generates, but also the losses that social obligations impose on the individual. In the calculus of optimal social values, as in the calculus of optimal taxation, both individual liberty and social obligations will have roles to play.[3]

The model rests on three main assumptions. The first assumption is that certain values and understandings are established at a level that is external to the individual. For example, the individual may belong to a nation, a religious congregation, a profession, a clan, a close family – each group endowed with some shared understandings and values. These understandings and values define the moral implications of the individual's behavior.

The second assumption is that individuals *internalize* the society's moral judgment. That is, the individual takes social understandings and values into account even in the absence

[2]The empirical literature on cultural differences in moral values and behavior is vast; see, for example, Henrich et al (2004), Falk et al (2018), and Inglehart (2018).

[3]We interpret Harrod (1936) as making essentially this point.

of external observers, rewards, or sanctions. In this sense, a particular passion – *guilt* – is involved in the production of decency. However, the internalization may be partial; the individual does not slavishly submit to the society's morality. In the model, the main source of heterogeneity is that the degree of decency varies across individuals.

The third assumption is that social understandings are incomplete. Reality is so vast that any workable rule necessarily depends only on a sparse description of the situation (Jehiel, 2005; Gabaix, 2014; Mohlin, 2014; Mailath, Morris, and Postlewaite, 2017). Therefore, not all actions that have desirable consequences are subject to moral judgment. For example, the individual might be supposed to help when confronting someone in need, but not to actively seek out the needy.[4]

A distinctive feature of our theory is that we explicitly consider the situation itself to be malleable. Socialization entails internalization of social values and understandings, and these sometimes interact in subtle ways.[5] For example, instilling concern for others' payoffs promotes cooperation in social dilemmas, but instilling awareness of others' choices can jeopardize cooperation.[6] Since mere awareness of others' making choices matters for the individual's choice, and since in lab experiments the complete interaction is typically described in detail and made common knowledge, the theory supports the view that lab-to-field generalization is challenging when social values and understandings matter for individual decision-making. Social context in the lab may easily fail to capture the social context in the field settings that the lab experiments intend to reflect (Levitt and List, 2007; List, 2009; Galizzi and Navarro-Martinez, in press).[7]

We define a social norm to be a strategy profile such that no player could raise social value through a unilateral deviation. It is shown that norms may mandate collectively optimal behavior, but need not do so. Actual behavior in turn coincides with some norm only if individuals are sufficiently decent. Seeking stronger predictions, we propose a structural model of social values; society endorses efficiency and equality, but condemns ill-gotten gains. The model was designed to capture the phenomenon that decent people tend to avoid situations that encourage pro-social behavior. We find it encouraging that it also

[4]One explanation for incomplete moral regulation is that it is much easier to identify clear moral failures under well-defined circumstances than to keep track of a person's accumulated morality. Laws likewise focus on defining and punishing specific instances of undesirable behavior.

[5]The relationship between objective social reality and subjective understanding of reality is also an old topic. In modern times, Berger and Luckman (1966) emphasize that social institutions require shared understandings of social reality, and they discuss the ways in which such understandings are developed by habituation and transmitted through socialization. Related themes are central to social psychology (e.g., Nisbett and Ross, 1991), where our model is particularly closely related to the interdependence theory of Kelley and Thibaut (1978); for an introduction see Rusbult and Van Lange 2008).

[6]On the one hand, awareness of others' choices increases individuals' desire to cooperate if others are expected to cooperate. On the other hand, it decreases individuals' desire to cooperate if others are expected to defect, and thus creates an opportunity for coordination failure.

[7]There are many instances of a close resemblance between lab and field behavior; see, for example, Östling et al (2011) and the references therein. But perhaps it is no coincidence that the greatest successes are from competitive situations in which concerns for decency are absent.

rationalizes several phenomena that it was not designed for, such as the existence of willful ignorance, intention-based negative reciprocity, and betrayal aversion.

Related literature. Decency has always played a central role in sociology. For example, both Emile Durkheim and Max Weber explicitly focus much of their analysis on internalized moral obligations.[8] Likewise, anthropologists have proposed that a central property of societies is their degree of cultural tightness.[9] Expressing such sociological ideas in the language of game theory allows us to naturally combine methodological individualism with group-level concepts such as shared values and understandings.

In mainstream economics, on the other hand, the role of decency has usually been implicit (Arrow, 1974), perhaps partly because passions and duties do not directly matter for general equilibrium analysis of frictionless markets (Dufwenberg et al, 2011). The neglect of decency has limited the reach of economic analysis, but should not be mistaken for a presumption of indecency. To the contrary, Friedman (1970), who often gets to epitomize the heartlessness of neoclassical economics, takes decency for granted:

> [The responsibility of a corporate executive] is to conduct the business in accordance with [owners'] desires, which generally will be to make as much money as possible while conforming to [the] basic rules of the society, both those embodied in law and those embodied in ethical custom.

Even Oliver Williamson, who develops a theory of economic organization that emphasizes the shortage of decency, does not assume that *all* people are "opportunistic with guile," but that *some* business people will be willing to lie and cheat for private profit (Williamson, 1975, p.26-27).

Behavioral economic theory takes on the task of modeling human moral motivation in more detail. The literature has hitherto emphasized the role of passions. Prosocial behavior has been modeled as altruism (Edgeworth, 1881; Becker, 1974), fair-mindedness (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), or taste for reciprocity (Rabin, 1993; Levine, 1998; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006).[10] However, behavioral economists always observed the complementary role of duty. Camerer and Thaler (1995) argue that behavior in Ultimatum and Dictator experiments is often better described in terms of "manners" rather than individual desires. Formal models of internalized social

[8] See especially Durkheim (1957/1900) and Weber (1930/1905).

[9] The concept of cultural tightness has a long history; see Pelto (1968) for a brief survey and a suggested definition. A much-cited empirical study is Gelfand et al (2010).

[10] Behavior can also be driven by desire for social esteem (e.g., Bernheim, 1994; Glazer and Konrad, 1996; Bénabou and Tirole, 2006; Ellingsen and Johannesson, 2008; Andreoni and Bernheim, 2009), but meaningful social esteem for prosocial traits requires that there are individual differences in these traits to begin with. In particular, Ellingsen and Johannesson (2011, Section 3) observe that social esteem concerns may spur generous behavior in circumstances where people would behave selfishly in the absence of esteem concerns, yet there would be no esteem effect without underlying differences in prosocial traits. A final theory of unselfish behavior that has been proposed to account for reluctant charity and willful ignorance relies on self-deception; see Bénabou and Tirole, 2011. We shall not here consider such departures from the standard model of beliefs.

norms include Bernheim (1994), Rabin (1994, 1995), Lindbeck, Nyberg, and Weibull (1999), Konow (2000), Bicchieri (2005), López-Pérez (2008), Krupka and Weber, (2013), and Spiekermann and Weiss (2016).[11] Akerlof and Kranton (2000) build a related model of social identity.[12]

We depart from these analyses of internalized social norms in two main ways. First, we establish a more basic foundation – a portable framework in which the norms themselves derive from social values and understandings. In this respect, our approach builds on Brekke, Kverndokk and Nyborg (2003). There is also a close formal similarity with Huck, Kübler and Weibull (2012), who assume that people maximize a combination of personal benefits and social value.[13] However, theirs is a model in which norms arise from passion rather than duty. This distinction is crucial once we relax the standard assumption that all situations are equally social. In particular, like Dillenberger and Sadowski (2012), we explicitly allow for the possibility that decision-makers can seek or avoid situations in which norms apply.[14]

An influential literature identifies social conventions (or descriptive norms) with equilibria (Lewis, 1969), in particular evolutionarily stable equilibria (Sugden, 1986), or stochastically stable equilibria (Young, 1992).[15] Our analysis of injunctive norms and social conventions likewise utilizes a refinement of Nash equilibrium, but does not consider the issue of evolutionary selection.

2 Model

The model formalizes conceptual linkages from social values to social norms as well as from social values to individual behavior.

Before introducing formal notation and definitions, let us provide a brief intuitive account of social situations and the moral preferences that we emphasize.

[11] The most closely related theory of social norms is probably that of Bicchieri (2005); we comment on the relationship below. Among the many other less formal approaches to social norms and related concepts, the spirit of our theory is close to Parsons (1951), Thibaut and Kelley (1959), Opp (1982), March and Olsen (1989,1994), and Coleman (1988,1990).

[12] Other formal approaches that apparently involve passion can potentially be re-interpreted in terms of decency. In particular, we think that Andreoni's (1989,1990) concept of impure (warm-glow) altruism is better understood as desire to fulfill duties than as "joy of giving."

[13] As will become clear, many of the applications that we have in mind also require different assumptions concerning the arguments and the shape of the value function.

[14] We assume that situation-avoidance does not involve any cognitive dissonance. By contrast, in the models of Rabin (1994) and Konow (2000) agents can relax the utility cost of norm violations by adjusting their personal definition of the norm at the cost of some cognitive dissonance. In another related contribution, Rabin (1995) models norms as a constraint on choice rather than as an element of the utility function. Consequently an agent wants to avoid or relax norms, much as a consumer would benefit from a relaxation of the budget constraint.

[15] Ullman-Margalit (1977) formulated an early game-theoretic account of social norms in three different classes of games. In coordination games and "partiality games" (e.g. Hawk-Dove games) her theory is that norms are selected equilibria (similar to Lewis) whereas in social dilemmas she identifies social norms with efficient non-equilibrium outcomes, requiring some kind of internalized social values.

2.1 A brief informal account

Social reality is complex. In order to navigate it, individuals and societies parse the vast web of interactions into manageable pieces. An individual's representation of such an excerpt of reality is called a "situation".[16]

Within a culture, some situations are considered to be social. In a social situation, individuals are supposed to pay attention to social values. To the extent that an individual fails to pay proper attention to social values, the individual is blameworthy and will suffer some guilt. Conversely, if the situation is not considered social, the individual may ignore social values without causing blame or guilt. Thus, the moral behavior that we consider is driven by internalized group-level understandings and objectives in general and by the avoidance of guilt in particular.

In unfamiliar situations, such as a laboratory experiment that does not immediately resemble a commonly recognized social situation, individuals may be uncertain about whether the situation is social or not.

2.2 Situations, games, and solution concepts

Apart from a slight generalization of the utility function, our basic definitions are standard.

Situations. A situation, or game form, is a tuple $F = \langle N, S, Z, x \rangle$, where $N = \{1, 2, \ldots, n\}$ is a finite set of players, $S = \times_i S_i$ is the finite set of pure strategy profiles, Z is a set of outcomes, and $x : S \to Z$ is an outcome function. For simplicity, we only consider material outcomes, so $Z \subset \mathbb{R}^n$ throughout.[17] Let $\Sigma = \times_i \Sigma_i$ denote the set of mixed strategy profiles, with σ being a typical element.

Games. In conventional game theory, Player i's preferences are captured by a von Neumann-Morgenstern utility function, $\tilde{U}_i : Z \to \mathbb{R}$, with $\tilde{u}_i(\sigma) := E_\sigma \left[\tilde{U}_i\left(x\left(s\right)\right)\right]$. Here, we shall allow richer (more sociological) preferences, while retaining the key property of expected utility theory that $u_i(\sigma) := E_\sigma \left[U_i\left(x\left(s\right)\right)\right]$. As usual, we let $u_i(s)$ denote the utility associated with the pure strategy profile s. A game is a tuple $G = \langle F, u \rangle$.

Solution concepts. The two solution concepts that we consider are Nash equilibrium and Undominated Nash Equilibrium.

Definition 1 *A strategy profile σ^* is a Nash equilibrium of a game G if, for all $i \in N$,*

$$\sigma_i^* \in \arg\max_{\sigma_i \in \Sigma_i} u_i(\sigma_i, \sigma_{-i}^*).$$

[16] Sociologists may recall the Thomas theorem: "If men define situations as real, they are real in their consequences" (Thomas and Thomas 1928).

[17] Among other things, this restriction prevents us from discussing morality in relation to communication. In order to study honesty, the space of strategies would need to include messages, a kind of action that does not map directly to material outcomes.

Definition 2 *A Nash equilibrium σ^* is undominated if there is no player i and strategy $\sigma_i \neq \sigma_i^*$ such that $u_i(\sigma_i, \sigma_{-i}) \geq u_i(\sigma_i^*, \sigma_{-i})$ for all σ_{-i} and $u_i(\sigma_i, \sigma_{-i}) > u_i(\sigma_i^*, \sigma_{-i})$ for some σ_{-i}.*

The solution concepts will be used for the purpose of generating predictions.[18] Specifically, when a game has an undominated equilibrium in pure strategies, we say that it is a potential convention (descriptive norm) of the associated situation.

Definition 3 *A strategy profile σ is a potential convention of the situation F if it is an undominated pure strategy Nash equilibrium of the game $\langle F, u \rangle$.*

Before using the model to make predictions, we introduce normative prescriptions, which will affect predictions to the extent that they are internalized by the agents.

2.3 Social values and social norms

Social values order the set of outcomes Z in social situations.[19] Thus, social values are expressed by a function $V : Z \to \mathbb{R}$. Let the social value associated with a strategy profile σ be denoted[20]

$$v(\sigma) := E_\sigma[V(x(s))]. \tag{1}$$

Social values are assumed to be constant across large classes of situations, thereby preventing the analyst from tailoring the value function to specific situations. This is the main source of the theory's predictive power.

Definition 4 *A social norm in a social situation F is a strategy profile σ^* such that, for all $i \in N$,*

$$\sigma_i^* \in \arg\max_{\sigma_i \in \Sigma_i} v(\sigma_i, \sigma_{-i}^*).$$

That is, in a social situation a social norm requests each player to pursue a strategy that maximizes social value (conditional on opponents' actions). Note the analogy with Nash equilibrium. Let $\Sigma^{PN}(F, v)$ denote the set of social norms in social situation F.

Our next definition singles out the norms that maximize social value.

Definition 5 *An ideal norm is a strategy profile*

$$\sigma^* \in \arg\max_{\sigma \in \Sigma} v(\sigma).$$

[18] Our reliance on Nash equilibrium to make predictions about behavior can be justified by the literature on evolution and learning in games. Typically the set of rest points of evolutionary dynamics contain the set of Nash equilibria, and sometimes these sets coincide. For an accessible recent introduction, see Young (2015), and for a comprehensive textbook treatment, see Sandholm (2010). Evolutionary motivations for eliminating weakly dominated strategies are more limited; see Nachbar (1990), Samuelson (1994), and Kuzmics (2011).

[19] Social values should not be confused with social welfare; see below.

[20] Later, we also consider value functions that depend on s other than through the final outcome x. We could also relax the assumption of social risk neutrality that is implied by the expectations operator; the analysis generalizes to other functions that vary continuously with σ.

That is, an ideal norm is a strategy profile that maximizes social value. We say that an ideal norm is *pure* if it prescribes a pure strategy profile. Since S is a finite set, $V(x(s))$ has a maximum. Let $\bar{S}(F,V) = \arg\max_s V(x(s))$ be the non-empty set of maximizers. Moreover, it follows from (1) that $\max_\sigma v(\sigma) = \max_s V(x(s))$. Our first result follows immediately.

Theorem 1 *For any social values V and situation F, there exists a non-empty set of pure ideal norms $\bar{S}(F,V) \subseteq \Sigma^{PN}(F,v)$.*

Given that social values are defined on final outcomes, it is intuitive that the set of norms includes all pure strategy profiles that maximize the social value function v.

As it turns out, there is often more than one (pure) strategy profile that maximizes social value v, especially in multi-stage situations. We therefore refine the set of norms as follows.

Definition 6 *A norm σ^* is* undominated *if there is no player i and strategy $\sigma_i \neq \sigma_i^*$ such that $v(\sigma_i, \sigma_{-i}) \geq v(\sigma_i^*, \sigma_{-i})$ for all σ_{-i} and $v(\sigma_i, \sigma_{-i}) > v(\sigma_i^*, \sigma_{-i})$ for some σ_{-i}.*

Let $\Sigma^{UPN}(F,v) \subseteq \Sigma^{PN}(F,v)$ denote the set of undominated norms in situation F.

Theorem 2 *For any values V and situation F, there exists a pure ideal norm $\bar{s} \in \bar{S}(F,V) \subseteq \Sigma^{UPN}(F,v)$.*

Proof. Suppose $\hat{s}$ is an ideal norm which is not undominated. Thus there is some i and some σ_i' that weakly dominates $\hat{s}_i$. Since $\hat{s}$ is an ideal norm, $v(\hat{s}_i, \hat{s}_{-i}) = v(\sigma_i', \hat{s}_{-i})$, meaning that $(\sigma_i', \hat{s}_{-i})$ is also an ideal norm. Let $C(\sigma_i')$ denote the support of σ_i'. Since $v(\sigma_i', \hat{s}_{-i}) = \sum_{s_i \in C(\sigma_i')} \sigma_i'(s_i)\, v(s_i, \hat{s}_{-i})$, the fact that v is maximized at $(\sigma_i', \hat{s}_{-i})$ implies that $v(s_i', \hat{s}_{-i}) = (\sigma_i', \hat{s}_{-i})$ for all $s_i' \in C(\sigma_i')$. Thus the profile $(s_i', \hat{s}_{-i})$ is an ideal norm for any $s_i' \in C(\sigma_i')$. Pick an $s_i' \in C(\sigma_i')$. If s_i' is undominated, then $(s_i', \hat{s}_{-i})$ is an ideal norm in which i plays an undominated strategy. If instead s_i' is weakly dominated then there is some σ_i'' that weakly dominates s_i' and $\hat{s}_i$. Iterate the argument to conclude that $(s_i'', \hat{s}_{-i})$ is an ideal norm for any $s_i'' \in C(\sigma_i'')$. Thus, either $(s_i'', \hat{s}_{-i})$ is an ideal norm in which i plays an undominated strategy, or there is some σ_i''' that weakly dominates s_i'', s_i', and $\hat{s}_i$. Since there are finitely many strategies, we eventually find an undominated strategy s_i^* such that $(s_i^*, \hat{s}_{-i})$ is an ideal norm. Iterating for each player i completes the construction of a pure ideal norm. ■

In other words, there is always a norm that tells each player exactly what to do, and that norm maximizes social value.[21]

There frequently exist additional norms that do not maximize social value. At first sight, such non-ideal norms appear unappealing. However, non-ideal norms are sometimes less demanding and can thus be easier to promote.[22]

[21] If the social value function had not been based merely on ex post outcomes, norms would not necessarily prescribe a certain action profile; randomization might then be preferable.

[22] A pragmatic norm might be defined as the norm that facilitates the best expected outcome conditional on the prevailing level of decency, where decency is defined precisely below.

Moreover, players may fail to obey any norm; social values may well affect behavior without determining it entirely. Let us now describe the way which players internalize the social values.

2.4 Blame, guilt, and utility

Society ascribes blame to players who fail to maximize social value in social situations. Let the blame $b_i : S \to \mathbb{R}$ assigned to Player i equal the social loss that Player i causes,

$$b_i\left(s_i, s_{-i}\right) := \max_{\bar{s}_i} V\left(x\left(\bar{s}_i, s_{-i}\right)\right) - V\left(x\left(s_i, s_{-i}\right)\right). \tag{2}$$

According to this definition, blame depends on what others are doing. For example, in a weak-link situation (such as Stag Hunt) no player is blamed for taking the lowest action if at least one other player does so – but a player is blamed for being the only player not to take the highest action.[23]

Players' concern for blame is captured by a normative utility component $U^b : \mathbb{R} \to \mathbb{R}$, whereas their concern for material payoff is captured by the material utility component $U^z : Z \to \mathbb{R}$. Observe that material utility is allowed to depend on the whole profile of material payoffs $x\left(s\right)$, not only the individual's own payoff $x_i\left(s\right)$. Thus we do not in general preclude passion (other-regarding preferences) to be part of motivation.

For simplicity, we assume that preferences are additively separable, with overall utility denoted

$$U_i\left(s\right) = U^z\left(x\left(s\right)\right) - \delta_i U^b(b_i\left(s\right)). \tag{3}$$

Define *guilt* as the disutility of blameworthiness, i.e., $\delta_i U^b(b_i\left(s\right))$. We refer to δ_i as Player i's degree of *decency*. Decency is the only source of preference heterogeneity that we consider.

Suppose players take blame into account even when nobody can observe their behavior. That is, players feel guilt when they are *blameworthy*. For example, guilt from blameworthiness may keep people from stealing in situations where they know that the crime could not

[23] One may consider a more deontologically flavored specification according to which blame depends not on what others do but on what they should do. If there is a unique pure ideal norm s^* (deontological) blame may plausibly be defined as

$$\tilde{b}_i\left(s_i, s_{-i} | s^*\right) := \max_{\bar{s}_i} V\left(x\left(\bar{s}_i, s^*_{-i}\right)\right) - V\left(x\left(s_i, s^*_{-i}\right)\right).$$

be discovered.[24] Finally, assume that Player i maximizes the expectation of U_i, that is,[25]

$$u_i(\sigma) := \quad E_\sigma \left[U_i \left(x_i \left(s \right), b_i \left(s \right) \right) \right] \quad = E_\sigma \left[U^z \left(x \left(s \right) \right) + \delta_i U^b \left(s \right) \right].$$

Note that our model differs from previous models of internalized social norms in that there is no cost of norm violation as such; no norm appears in (3). Instead, players merely feels guilty about not maximizing social value. As it happens, however, the magnitude of guilt will often endogenously acquire characteristics that previous authors, such as Bicchieri (2005) and López-Pérez (2008), have assumed. For example, in many situations (but not all) the lost social value associated with an individual's deviation from a norm will be greater when others do not similarly deviate. Thus, it is *as if* the individual feels more compelled to comply with the norm if others also comply.[26]

2.5 A simple linear model

In order to derive sharp predictions from the model, we must make additional assumptions about the social value function. In general, appropriate assumptions depend on the nature of the situation as well as on the particulars of the culture under consideration.

For purposes of illustration, let us initially focus on the simple social value function

$$V \left(x \right) = x^+ - \alpha x^-, \tag{4}$$

where

$$\begin{aligned} x^+ &:= \sum_{i=1}^{n} x_i, \\ x^- &:= \sum_{i,j:i \neq j} \max \left\{ 0, x_i - x_j \right\}. \end{aligned}$$

[24] Our concept of guilt is broadly in line with standard psychological definitions. For example, Haidt (2003) writes: "As the traditionally central moral emotion, guilt was said to be caused by the violation of moral rules and imperatives [...], particularly if those violations caused harm or suffering to others [...]. The literature on morality has many other names for the passions that sustain obligations. For example, Gouge (1622) used both the concepts of *conscience* and *filial fear* (as opposed to slavish/servile fear) for this passion, as noted by Kahn (1999). These traditional concepts of guilt from causing harm are related to but different from the recent concept of guilt defined by Charness and Dufwenberg (2006). According to their definition, people experience guilt when they disappoint others. Here, it is not others' disappointment or disapproval as such that matters, but whether the action would have qualified for disapproval if others had known it.

[25] We do not offer an axiomatic foundation for this representation, but see Dillenberger and Sadowski (2012) and Breitmoser and Vorjohann (2017) for related efforts in that direction.

[26] Bicchieri (2005) defines a social norm as a behavioral rule for a class of situations, such that, for each member of the community, (i) the player knows that the rule exists and applies to the class of situation, and (ii) prefers to comply with the rule provided that (a) the player believes that others will comply and (b) the player believes that others thinks that she ought to comply. Since Bicchieri's definition does not link the norms to the prevailing social values, the pressure to comply with a norm is driven by others' norm compliance rather than by the social losses caused by non-compliance.

That is, society puts positive value on efficiency (x^+) and negative value on inequality (x^-). (Later, we shall add to this social value function a dislike for ill-gotten gains. However, as this feature is irrelevant to our initial applications, we ignore it for now.)

For simplicity, we also assume that utility functions are linear.

$$u_i(s) = x_i(s) - \delta_i b_i(s). \tag{5}$$

In the Appendix we consider a non-linear specification where the cost of blame includes both a fixed component and a variable convex component (c.f. Abeler, Nosenzo, and Raymond, 2016; Malmendier, Velde, and Weber, 2014).

3 Socialization and social dilemmas

Our first examples illustrate the various components of socialization. We consider a society in which row players meet column players. The pairings are new each time, and players have no knowledge of the opponent's previous history.

The material payoffs associated with a row player's action are depicted in the left panel of Figure 1, and the material payoffs associated with the column player's action are depicted in the right panel. As usual, the first number in each cell is the row player's payoff and the second number is the column player's payoff. The interaction is a Social Dilemma, because the action that maximizes the decision maker's payoff fails to maximize overall payoff. By taking action C (cooperate) rather than action D (defect), a player would make a sacrifice for the benefit of the other.

Note that we represent the payoffs in separate bi-matrices; call this representation narrow social bracketing.

(C)	$0, (3)$
D	$1, (0)$

Rowena's choice

(C)	D
$(3), 0$	$(0), 1$

Colin's choice

Figure 1: The narrowly bracketed Social Dilemma

Consider a row player, Rowena. Suppose she is born a selfish materialist and receives no socialization. Thus, her utility equals her material payoff, say $u_R = x_R$, and she does not pay attention to Colin's payoffs. Therefore these payoffs are in brackets in the figure. Since she does not pay attention to Colin's payoffs, there is also no reason to pay attention to the action C itself. Indeed, helping Colin by playing C might not seem much more relevant to Rowena than, for example, kicking him or giving him some money – feasible actions that even we, the analysts, does not consider. The logic for a column player, say Colin, is the same.

Proposition 1 *In the narrowly bracketed Social Dilemma, the unique potential convention is* (D, D).

So far, this is nothing but an additive version of the Prisoners' Dilemma.

We model socialization as an expansion of the players' mental accounting. Socialization can be divided into two separate steps.

3.1 Contemplating others' interests

The first step is to make the situation social. In other words, let Rowena be informed that her choice has consequences for Colin, and that the society's values instruct her to take Colin's outcomes into account. To be precise, Rowena is instructed to consider the social values $v = x^+ - \alpha x^-$ in addition to her own material payoff x_R; Figure 2 illustrates this dutifully empathic ideal in such a narrowly bracketed social dilemma.

C	$3 - 3\alpha$
D	$1 - \alpha$

Figure 2: Rowena's dutifully empathic ideal

Comparing the two alternatives yields the norm.

Proposition 2 *In the narrowly bracketed Social Dilemma, the norm is* C *if* $\alpha < 1$ *and* D *if* $\alpha > 1$.

The intuition is plain. Rowena should help Colin if and only if payoff-maximization is sufficiently important relative to distributional concerns.

Of course, Rowena will only comply with this norm if her decency is sufficiently great. Suppose $\alpha < 1$. Rowena's actual decision problem after partial socialization is then depicted in Figure 3.[27]

C	0
D	$1 - (2 - 2\alpha)\delta_R$

Figure 3: Rowena's partially socialized game

Proposition 3 *Suppose players are equally decent, and suppose* $\alpha < 1$. *Then, the unique potential convention in the narrowly bracketed Social Dilemma is* (C, C) *if* $\delta > 1/(2 - 2\alpha)$ *and* (D, D) *if* $\delta < 1/(2 - 2\alpha)$. *If* $\alpha > 1$ *the unique potential convention in the narrowly bracketed Social Dilemma is* (D, D).

That is, there is a potential convention of cooperation if players are sufficiently decent.

[27] Note that $b = 3 - 3\alpha - (1 - \alpha) = 2(1 - \alpha)$.

3.2 Contemplating others' behavior

The second component of socialization is the expansion of mental accounting to recognize others as agents and consider their similar motives. That is, Rowena is instructed to think about Colin's choice as well as her own. We call this "broad social bracketing." She ought to pay attention, at least initially before behavior has settled down, to four strategy profiles $\{(C,C),(C,D),(D,C),(D,D)\}$. These strategy profiles have associated material payoff profiles as in the upper left panel of Figure 4 and social payoff profiles as in the upper right panel of in Figure 4. Suppose $\alpha \geq 1/2$. Then, the lower panel is the game as played by two broadly bracketing players.

	C	D
C	$3,3$	$0,4$
D	$4,0$	$1,1$

Material payoffs

	C	D
C	6	$4(1-\alpha)$
D	$4(1-\alpha)$	2

Social payoffs

	C	D
C	$3,3$	$-\delta_R(4\alpha-2), 4-\delta_C(2+4\alpha)$
D	$4-\delta_R(2+4\alpha), -\delta_C(4\alpha-2)$	$1,1$

The fully socialized Social Dilemma game if $\alpha \geq 1/2$

Figure 4: Broad social brackets

Under broad social bracketing, (C,C) is a norm regardless of α, since this strategy profile simultaneously maximizes x^+ and minimizes x^-. However, the profile (D,D) is also a norm. If Colin is expected to play D, and society values equal payoffs, (i.e., $\alpha \geq 1/2$), Rowena contributes more to attaining social values by playing D than by playing C. Intuitively, (D,D) can be attractive even when D alone is not, because D alone is associated with some inequality, but (D,D) is not.

Does more complete socialization encourage more cooperation? The answer is yes and no.

Proposition 4 *Suppose players are equally decent and that they bracket the Social Dilemma broadly. Then (D,D) is always a potential convention, whereas (C,C) is a potential convention if and only if $\delta \geq 1/(2+4\alpha)$.*

On the one hand, the minimal decency $\underline{\delta}$ that is compatible with a cooperative convention is smaller under broad social bracketing, and strictly so if $\alpha > 0$. The intuitive reason is that the inequality generated by unilateral play of D is doubled, because the defecting player takes into account the fact that the opponent cooperated. On the other hand, under narrow social bracketing, cooperation is guaranteed whenever decency is above the relevant threshold. Under broad social bracketing, the uncooperative outcome (D,D) is a strict

equilibrium regardless of δ and α.[28]

These computations depend on a homogeneous level of decency. With observable heterogeneous decency, broad social bracketing has the feature that cooperation breaks down completely whenever one player's decency is below the threshold. If social values are little concerned with equality (α is low), narrow social bracketing would then gain an additional advantage over broad social bracketing. Arguably, the most realistic case involves unobservable heterogeneous decency. In this case, the relevant solution concept is Bayesian equilibrium. All decency-types playing D remains an equilibrium. If the type distribution contains a sufficiently large fraction of highly decent players, there will also be equilibria in which players above a certain decency-threshold play C. Qualitatively, the comparisons remain as in the case of complete information, but quantitatively the requirements for sustaining C will differ.

To the extent that the construction of social situations serves the purpose of supporting cooperation, the analysis suggests that broad social bracketing will go together with high egalitarianism and widespread decency.

3.3 Bracketing alone

While our present focus is on social values and decency, we should note that there are many cases in which the scope of the brackets alone is important. For example, suppose players could observe each others' history. Then, the potential strategy space at time t is all mappings from the completed history to today's action. It is well understood that patient players can in principle sustain cooperation with the help of history-dependent strategies (even if the same pair of players never meets twice; see Kandori, 1992). However, it is far from clear that players would be able to experiment their way to a cooperative outcome. Socialization might here have a key role in helping selfish players to coordinate on cooperative equilibria by suitably adapting the social bracketing. In games with indefinite horizon, and other games with multiple equilibria, the role of morality is to help people coordinate; see Binmore (2005).

3.4 Reciprocity: A first take

Consider again the same actions of Figure 1, but now assume that Colin observes and recalls Rowena's action before taking his own. Suppose bracketing is broad. Then, the strategies and material payoffs are as given in Figure 5.

Here, the strategy CC means that Colin plays C regardless of what Rowena has played, whereas the strategy CD means that Colin plays C if Rowena played C and D if she played

[28] If one believes that risk-dominant equilibria are selected in the long run, as implied by Foster and Young (1990), Young (1993), and Kandori, Mailath and Rob (1993), this means that broad social bracketing is only preferable to narrow social bracketing if $\delta > 1/(1+2\alpha)$.

	CC	CD	DC	DD
C	$3,3$	$3,3$	$0,4$	$0,4$
D	$4,0$	$1,1$	$4,0$	$1,1$

Figure 5: Sequential Social Dilemma situation, broad brackets

D, and so on. Suppose $\alpha > 1/2$. Material payoffs translate into social value as before, and the norms are $(C, CC), (C, CD)$, and (D, DD). However, CC and DD are both weakly dominated by CD (since $\alpha > 1/2$ is equivalent to $2 > 4(1-\alpha)$). Thus, $(C, CD)$is the only undominated norm.

Cooperation is now also more likely to be the unique potential convention.

Proposition 5 *Assume $\alpha > 1/2$. Suppose players bracket the sequential Social Dilemma broadly. If $\delta_C \geq 1/(2+4\alpha)$, then (C, CD) is the unique potential convention.*

Two things are noteworthy about this result. First, the risk of coordination failure disappears, because of sequential moves. Second, the first mover does not need to be decent at all in order to cooperate.

In this case of sequential moves, it is straightforward to allow for incomplete information about δ. Let us compute the unique Perfect Bayesian equilibrium. Colin will always respond to D by playing D, and will respond to C by C if $\delta_C \geq 1/(2+4\alpha)$ (same threshold as above). Rowena's strategy depends on the type distribution that Colin is drawn from. A sufficient condition for Rowena to play C independently of her own type is that this is the optimal strategy when $\delta_R = 0$. In this case, $u_R(D) = 1$ and $E[u_R(C)] = 3 \cdot Pr(\delta_C \geq 1/(2+4\alpha))$. Thus, an even an indecent Rowena plays C as long as there is a probability of at least $1/3$ that Colin plays CD.

Note that the reciprocity norm arises as a consequence of the specific social values, putting enough weight on equality. If $\alpha < 1/2$, the unique undominated norm is instead unconditional cooperation, (C, CC). (Sometimes, reciprocity is defined in terms of intentions rather than mere actions; we return to this issue in Section 5.)

The model is able to fit many patterns in social dilemma experiments (for a recent summary of relevant findings, see Fehr and Schurtenberger, 2018), in particular it is consistent with the large prevalence of conditional cooperation, and the fact that cooperation seems much more sensitive to minor contextual changes when moves are simultaneous than when they are sequential – a marker for multiple equilibria in the simultaneous setting. However, as noted by Ellingsen et al (2012), these findings are all consistent with standard social preference models as well.

So far, the novel predictions are therefore confined to more sociological hypotheses, for example that broad bracketing will go together with high egalitarianism and wide-spread decency.

4 Dictator situations

The Dictator situation involves two players, Player 1 and Player 2, with strategy sets $S_1 = \{0, 1, \ldots, 10\}, S_2 = \{\emptyset\}$. Material payoffs are $x_1 = s_1$ and $x_2 = 10 - s_1$.

All allocations are efficient. Hence, the unique value-maximizing allocation, and thus the only social norm, is the equal split, $s_1 = 5$.

According to the model, which behaviors would we expect to see? The dictator maximizes $U_1 = x_1 - \delta_1 b_1$. Expressing this in terms of s_1, we have

$$u_1 = s_1 - \delta_1 \cdot \alpha \left| s_1 - (10 - s_1) \right|.$$

For $s_1 < 5$, utility is increasing in s_1. For $s_1 \geq 5$, utility is increasing in s_1 if and only if $\delta \leq 1/2\alpha$. The answer follows.

Proposition 6 *In the Dictator situation, the amount kept is*

$$s_1 = \begin{cases} 10 & \text{if } \delta_1 \leq \frac{1}{2\alpha}; \\ 5 & \text{otherwise.} \end{cases}$$

This simple model accounts for most of the observed behavior in Dictator experiments, but misses the significant fraction of offers strictly between 5 and 10.[29]

4.1 Avoiding Other Players

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, dictators (Player 1) are told that recipients (Player 2) are not yet aware of the experiment. Player 1 is given the option to exit for a price of 1. In the case of exit, Player 1 thus keeps 9, and Player 2 will never be informed. The puzzle is that a significant fraction of subjects choose to exit.[30]

Suppose Player 1 views the whole Dictator experiment with exit option as a single social situation.[31] After the exit option is presented, she faces the choice set $\tilde{S}_1 = \{e, s_1\}$, where e denotes exit, and s_1 is the choice she made before the exit option was revealed. Since the whole experiment is viewed as a single social situation, the exit choice is viewed as a choice *within* a social situation and hence it is subject to social values. We assume that the

[29] In the Appendix, we consider a generalization that admits these choices. The blame function b has both a fixed component associated with all deviations from the value optimum and a variable component that is strictly convex rather than linear. For a closely related analysis, see Malmendier, Velde, and Weber (2014).

[30] This finding has been extensively replicated and elaborated. For laboratory experiments, see Broberg, Ellingsen, and Johannesson (2007) and Lazear, Malmendier, and Weber (2012); for field experiments, see DellaVigna, List, and Malmendier (2012) and see Andreoni, Rao, and Trachtman (2017), and for an observational study, see Knutsson, Martinsson, and Wollbrant (2013).

[31] This subsection has much in common with Malmendier, Velde, and Weber (2014).

consequences of choosing e or s_1 are evaluated from the perspective of what could have been obtained in the experiment as a whole.

Proposition 7 *Suppose the exit choice is seen as a choice* within *a social situation. Then* s_1 *dominates* e *irrespective of* δ_1

Proof. The utility associated with s_1 is $u_1(s_1) = s_1 - \alpha\delta_1 |2s_1 - 10|$ and the utility associated with e is $u_1(e) = 9 - \alpha\delta_1(9-0)$. We know that the optimal s_1 is either 10 or 5. First suppose $\delta_1 \leq 1/2\alpha$ so that $s_1 = 10$. In this case $u_1(s_1) = u_1(10) = 10 - 10\alpha\delta_1 > 9 - 9\alpha\delta_1 = u_1(e)$ if and only if $\delta_1 < 1/\alpha$, which is implied by $\delta_1 \leq 1/2\alpha$. Next suppose $\delta_1 > 1/2\alpha$ so that $s_1 = 5$. In this case $u_1(s_1) = u_1(5) = 5 > 9 - 9\alpha\delta_1 = u_1(e)$ if and only if $\delta_1 > 4/9\alpha$, which is implied by $\delta_1 > 1/2\alpha$. ■

So, why do subjects exit? The model offers a straightforward resolution. The original Dictator game is a canonical distribution task for which society should have developed a common understanding, hence subjects treat it as a social situation. However, the exit option creates uncertainty about the situation, so that at least a subset of the subjects do not consider the exit decision as part of a social situation. Consequently their exit decision is not itself subject to social blame.[32]

Thus, the original utility $u_1(s_1) = s_1 - \delta_1\alpha |2s_1 - 10|$ should be compared with $u_1(e) = 9$ rather than with $u_1(e) = 9 - 9\alpha\delta_1$.

Proposition 8 *Suppose the exit choice is seen as a choice between social situations, not within. Then,*

$$\tilde{s}_1 = \begin{cases} e & \text{if } \delta_1 \geq \frac{1}{10\alpha}; \\ s_1 & \text{otherwise.} \end{cases}$$

Proof. First suppose $\delta_1 \leq 1/2\alpha$ so that $s_1 = 10$. In this case $u_1(s_1) = u_1(10) = 10 - 10\alpha\delta_1 > 9 = u_1(e)$ if and only if $\delta_1 < 1/10\alpha$ (implying $\delta_1 \leq 1/2\alpha$). Next suppose $\delta_1 > 1/2\alpha$ so that $s_1 = 5$. In this case $u_1(s_1) = u_1(5) = 5 < 9 = u_1(e)$ always. ■

In other words, the most decent player types exit, whereas the least decent of those who initially keep the full amount prefer to abide by their original decision. Qualitatively, the prediction is in line with the data, which indicate that subjects are more likely to exit the less they had been keeping; see Lazear, Malmendier, and Weber (2012).

With this formulation, the puzzle is not that some subjects exit, but that some subjects who gave positive amounts refrain from exiting. Perhaps the most natural explanation is that some subjects could not bring themselves to see the exit as cancelling the moral obligations they had been confronted with. It is as if the exit decision is part of a social situation.

[32] In this way our explanation builds on the suggestion by Lazear, Malmendier, and Weber (2012), that subjects in the role of Player 1 consider the choice to be between (i) remaining in a situation involving both themselves and a recipient and (ii) a situation that involves only themselves.

The model likewise captures the findings of the field experiments of DellaVigna, List, and Malmendier (2012) and Andreoni, Rao, and Trachtman (2017). In both experiments, many people are revealed to be systematically avoiding the solicitation of charitable contributions. Apparently, some people avoid solicitors because they know they would be giving, whereas others avoid solicitors in order not to feel the guilt that is associated with not giving.[33]

4.2 Avoiding Payoff Information

Dana, Weber, and Kuang (2007) (DWK) conduct another intriguing experiment. Player 1 chooses between two actions, A and B, which determine own payoffs as well as the payoffs of Player 2. However, Player 1 is unsure of the situation.

	State 1 (Non-aligned)	State 2 (Aligned)
A	$6, 1$	$6, 5$
B	$5, 5$	$5, 1$

Figure 6: Payoffs in DWK

Action A always gains one unit of material payoff to Player 1. In State 1 (non-aligned), Player 1's gain comes at a loss of 4 to the opponent. In State 2 (aligned), the opponent instead avoids a loss of 4 when Player 1 takes the self-interested action A; see Figure 6. Let p denote the probability of the aligned State 2. In the benchmark treatment, $p = 1/2$. Before making the choice, Player 1 has the opportunity to learn the state for free.

Suppose Player 1 thinks about all the possible actions as belonging to one social situation; call this situation Full. Let R and N denote "revealing state" and "not revealing state" respectively. The strategy set of Full comprises six strategies. Let RAA denote "reveal and take action A in both states;" RAB denote "reveal and take action A in state 1 and B in state 2;" and NA denote "not reveal and take action A." Thus, the six strategies are $\{NA, NB, RAA, RAB, RBA, RBB\}$. Figure 7 summarizes the material payoffs to Player 1.

Strategy	$E[x_1]$	$E[\hat{V}]$
NA	6	$7 + 4p - \alpha\,(1 + 4\,(1 - p))$
NB	5	$10 - 4p - 4\alpha p$
RAA	6	$7 + 4p - \alpha\,(1 + 4\,(1 - p))$
RAB	$6 - p$	$6 + (1 - p) - \alpha\,(4 + (1 - p))$
RBA	$5 + p$	$10 + p - \alpha p$
RBB	5	$10 - 4p - 4\alpha p$

Figure 7: Strategies and Payoffs in Full

[33] Of course, if society deemed that people are already part of a social situation when they decide whether to avoid the solicitation or not, this logic would not work. However, we think that there are limits to how broadly situations might productively be bracketed. We comment briefly on this below.

The value-maximizing strategy in Full is RBA for all $\alpha > 0$, since this strategy implements both maximal total payoff and minimal inequality. The three strategies $\{RAB, RBB, NB\}$ are dominated for all Player 1 types, and the two strategies RAA and NA are payoff-equivalent. Thus, Player 1's choice is effectively between being selfish and always playing A or being unselfish by revealing and then playing BA. The selfish choice produces expected utility

$$u(NA) = u(RAA) = 6 - \delta_1(1-p)(3+5\alpha),$$

and the unselfish choice produces expected utility

$$u(RBA) = 5 + p.$$

The result follows.

Proposition 9 *Suppose the revelation choice is seen as a choice within the social situation Full. Then, Player 1 chooses the value-maximizing strategy RBA if $\delta_1 > 1/(3+10\alpha p)$ and either RAA or NA if $\delta_1 < 1/(3+10\alpha p)$.*

Suppose instead that Player 1 thinks about the revelation decision as a choice *between* social situations. If Player 1 reveals the state, she is either in situation Aligned or in situation Non-aligned. If Player 1 does not reveal, we say she is in situation Unknown.

Proposition 10 *Suppose the revelation choice is seen as a choice between social situations, not within. Player 1 chooses not to reveal and to take action A regardless of her decency δ_1 if*

$$p > \frac{3+5\alpha}{8(1+\alpha)}. \tag{6}$$

Proof. If Player 1 finds herself in Aligned, the choice is trivial. Action A is dominant regardless of Player 1's decency. To see this formally, compute the value associated with each of the two actions in Aligned: $v(A) = 6+5-\alpha(6-5) > v(B) = 5+1-\alpha(5-1)$. Thus, there is no blame associated with action A and positive blame associated with action B. Comparing $u_1(A)$ to $u_1(B)$, we have $u_1(A) = 6 > u_1(B) = 5 - \delta_1(v(A) - v(B)) = 5 - \delta_1(5+3\alpha)$.

If Player 1 finds herself in Non-aligned, the choice is more complicated. Now, $v(B) = 5+5 > v(A) = 6+1-\alpha(6-1)$, so the self-serving action A is subject to blame, while B is not. Accordingly, $u_1(B) = 5$ and $u_1(A) = 6 - \delta_1(3+5\alpha)$. It follows that Player 1 takes action A if $\delta_1 < 1/(3+5\alpha)$ and B otherwise.

If Player 1 finds herself in Unknown, whether the choice is conflicted or not depends on the parameters. Note that $v(NA) - v(NB) = 7 + 4p - \alpha(1 + 4(1-p)) - (10 - 4p - 4\alpha p)$. Thus, in Unknown, taking action A comes with no blame as long as (6) holds; $u(NA) = 6$. Revealing and subsequently taking the value-maximizing action yields utility $u(RBA) = 5 + p$. Revealing and subsequently taking action A yields $u(RAA) = 6 - \delta_1(5+3\alpha)$. Thus, N dominates R as claimed. ■

In DWK's experiment, half the subjects chose not to reveal, and to take action A. Once we consider Unknown as a social situation, the puzzle is not why so many subjects made that choice – it is the prediction of Proposition 10 for $\alpha < 1$ – but why so many did not.

This experiment has been both replicated and modified. Feiler (2014) varies p and finds that reductions in p are associated with significant increases in revelation. With our value function, this is natural. Action A is no longer the morally superior choice (for any $\alpha > 0$) if $p < 3/8$. As p drops to low levels, it is unattractive to remain in situation Unknown. Unknown looks increasingly similar to situation Non-aligned, and subjects will feel pressure to take action B. Once a subject prefers to take action B in situation Unknown, reveal is a better option.[34]

Grossman (2014) makes the observation that DWK effectively treats the situation Unknown as a default. He compares the original design to a design without any default, and finds significantly more revelation. Explaining Grossman's finding goes to the heart of our distinction between the choice situation of a person and the social situation. The social situation is defined at the group level, and is hence external to the individual. In laboratory experiments, the researcher can influence what the social situation is. Dana, Weber, and Kuang effectively defined Unknown to be a social situation. The subject can potentially move to another situation with less uncertainty by pressing a button, but this is more of a private choice, at least this is how a substantial fraction of subjects appear to perceive it. Grossman constructs an alternative social situation in which one choice is to become informed and another choice is to remain uninformed. Since there is no default, the choice whether to become more informed is an integral part of the social situation, and hence of a morally relevant choice.

Bartling, Engl, and Weber (2014) experimentally study the reaction of third-party observers of willful ignorance. The observers can engage in costly punishment of Player 1. There are two main findings. On the one hand, willfully ignorant dictators are punished less if their actions lead to unfair outcomes than dictators who reveal the consequences before implementing the same outcome. On the other hand, willfully ignorant dictators are punished more than revealing dictators if their actions lead to fair outcomes. The first finding is in line with the interpretation that situation Unknown is recognized as a social situation. Ignorance is at least to some extent a valid excuse. The second finding again suggests that this interpretation is not universal; some people view Full as the relevant social situation, and for them ignorance is immoral.

[34] In a related experiment, van der Weele (2014) varies the payoffs to the two parties, finding that revelation is affected more by the decision-maker's payoff than the opponent's payoff.

5 Intentions and reciprocity

Decency is potentially also involved in reciprocal actions – actions that purposefully reward and punish others' behavior. As is well understood, such reciprocity is tightly linked to notions of intentions. People are more prone to punish intentionally selfish acts and to reward intentionally unselfish acts, than acts which merely happen to promote selfish or unselfish causes.

Previous theories of reciprocity have mainly extended the game-theoretic framework either by allowing players to care about their opponent' preferences (e.g., Levine, 1998; Segal and Sobel, 2007; Ellingsen and Johannesson, 2008) or their opponents' beliefs (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). Here, we do not need to extend game theory in those ways. It suffices to maintain our previous extension that players take social values into account. Specifically, we now assume that society has a distaste for (intentionally) ill-gotten gains.[35]

5.1 A model of ill-gotten gains

Player i's ill-gotten gains are those additional material payoffs that she obtains when she fails to take an available action that would maximize social value. Formally, when Player i's opponents play s_{-i}, Player i maximizes social value by playing an action in the set of V-best responses,

$$S_i^* (s_{-i}) := \arg\max_{\tilde{s}_i} V (\tilde{s}_i, s_{-i}) .$$

Thus, Player i's ill-gotten gains are

$$G_i (s) := \max \left\{ 0, x_i (s_i, s_{-i}) - \max_{\tilde{s}_i \in S_i^*(s_{-i})} x_i (\tilde{s}_i, s_{-i}) \right\},$$

and aggregate ill-gotten gains are

$$G^+ (s) := \sum_{i=1}^{n} G_i (s) .$$

Accordingly, the social value function becomes

$$\hat{V} (x, s) = V (x) - \mu G^+ (s) = x^+ - \alpha x^- - \mu G^+ (s) . \tag{7}$$

[35] In this respect our approach resembles the models of reciprocity by Cox, Friedman, and Gjerstad (2007) and Malmendier and Schmidt (2017), who model reciprocity as a change in the weight put on another player's payoff in response to the payoff consequences of previous actions by that player.

Call V the *core value function* and $\hat{V}$ the *extended value function*. Norms are defined as before, but with $\hat{V}$ in place of V. Let $g^{+}(\sigma) := E_{\sigma}[G^{+}(s)]$ and

$$\hat{v}(\sigma) := E_{\sigma}\left[\hat{V}(x(s), s)\right] = v(\sigma) - \mu g^{+}(\sigma).$$

As before, a norm is a strategy profile σ^{*} such that, for all $i \in N$,

$$\sigma_{i}^{*} \in \arg\max_{\sigma_{i}} \hat{v}\left(\sigma_{i}, \sigma_{-i}^{*}\right),$$

and blame takes the form

$$\hat{b}_{i}(s_{i}, s_{-i}) := \max_{\bar{s}_{i}} \hat{V}(x(\bar{s}_{i}, s_{-i})) - \hat{V}(x(s_{i}, s_{-i})). \tag{8}$$

Accordingly, the utility function is

$$U_{i}(s) = U^{z}(x(s)) - \delta_{i} U^{\hat{b}}(\hat{b}_{i}(s)).$$

(A richer model would include a separate decency parameter associated with ill-gotten gains, breaking the tight link between negative and positive reciprocity.)

5.2 Numerical values

In our examples, we shall assume that the aversion to inequality is moderate, with $\alpha < 1/2$. Thus, in a two-player situation, the social value V goes up if there is an increase in the material payoff of one player while the material payoff of the other is constant.

On the other hand, we assume that the social aversion to ill-gotten gains is relatively strong, with $\mu > 10/3$. That is, when a player obtains one additional unit of material payoff by deviating from socially desirable behavior, the social value shrinks by more than thrice that amount.

Finally, we assume that Player i's decency δ_{i} is distributed on some interval $[0, \hat{\delta}]$ where $\hat{\delta}$ is positive and finite. Let D denote the cumulative distribution function; D has no subindex i, as we assume that all players are drawn from the same distribution. This specification is quite unrestrictive, as it only rules out negative decency.

5.3 An Ultimatum situation

Consider the binary Ultimatum situation (Figure 8). Player 1 first chooses either $s_{1} = F$, which ends the situation with even payoffs (5,5), or $s_{1} = U$ which continues the situation. In the latter case, Player 2 has the choice between $s_{2} = A$, which yields payoffs (8,2), or $s_{2} = P$, which yields (0,0).

Note that our general parameter assumptions imply $0 < 10 - 6\alpha < 3\mu$. Computing

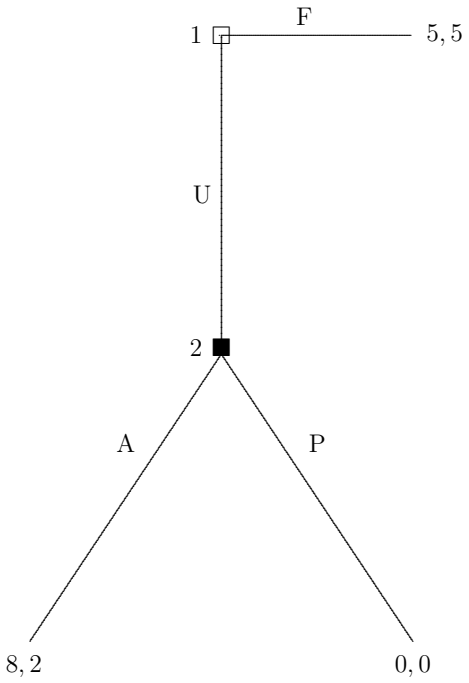

Figure 8: Ultimatum situation

the core value function V yields $V(F,A) = V(F,P) = 10$, $V(U,A) = 10 - 6\alpha$, $V(U,P) = 0$, implying that both (F,A) and (F,P) are V-norms. The extended value function is $\hat{V}(F,A) = \hat{V}(F,P) = 10$, $\hat{V}(U,A) = 10 - 6\alpha - 3\mu$, $\hat{V}(U,P) = 0$. It follows immediately that the undominated $\hat{V}$-norm impels Player 2 to play P rather than A.

Proposition 11 *In the Ultimatum situation, the unique norm is* (F,P).

Proof. The profile (F,P) dominates the the profile (F,A) because they yield the same $\hat{V}$ but $\hat{V}(U,P) > \hat{V}(U,A)$. ■

Having derived the extended value function, the players' utilities are as in Figure 9. Let

	A	P
F	$5\ ,\ 5$	$5\ ,\ 5$
U	$8 - \delta_1(6\alpha + 3\mu)\ ,\ 2 - \delta_2(6\alpha + 3\mu - 10)$	$-10\delta_1\ ,\ 0$

Figure 9: Ultimatum game

us now characterize behavior in the Ultimatum game.

Proposition 12 *In the unique Bayesian Nash equilibrium of the Ultimatum game, (i) Player 2 plays* P *if and only if*

$$\delta_2 \geq \delta^* = \frac{2}{6\alpha + 3\mu - 10};$$

(ii) Player 1 plays F if and only if

$$\delta_1 \geq \frac{8D\left(\delta^*\right) - 5}{(6\alpha + 3\mu)D\left(\delta^*\right) + 10\left(1 - D\left(\delta^*\right)\right)}.$$

Proof. Player 2's payoff is unaffected by Player 2's action if Player 1 plays F. Player 2's best response to U is to play P if and only if $0 \geq 2 - \delta_2(6\alpha + 3\mu - 10)$, or equivalently $\delta_2 \geq 2/(6\alpha + 3\mu - 10)$, proving part (i). Player 1's best response to this strategy by Player 2 is to play F if and only if

$$5 \geq D\left(\delta^*\right)\left(8 - \delta_1\left(6\alpha + 3\mu\right)\right) + \left(1 - D\left(\delta^*\right)\right)\left(-10\delta_1\right).$$

Solving for δ_1 yields (ii). ■

Part (i) says that a sufficiently decent Player 2 will punish, whereas any less decent Player 2 will not do so. Thus, we might expect some heterogeneity in Player 2 behavior. If the probability that Player 2 punishes, $1 - D(\delta_2^*)$, exceeds 3/8, we see from condition (ii) that Player 1 plays F regardless of the own decency δ_1 (which is never negative). Otherwise, only a sufficiently decent Player 1 plays F.

Since the model rationalizes punishment, we go on to investigate whether it does so for the right reason. Blount (1995) conducts a revealing experiment. In addition to a standard Ultimatum treatment, she considers how Player 2 behaves when Player 1's action is beyond Player 1's control.[36] Specifically, Player 1's action is picked by a computer programmed by the experimenter. Call this the Involuntary Ultimatum situation. Blount finds that punishment is sharply reduced in the Involuntary Ultimatum situation. According to our model, this is understandable. Punishment ceases to be a norm, because there are no ill-gotten gains when Player 1 could not choose F.

Proposition 13 *In the Involuntary Ultimatum situation, action A by Player 2 is both the unique undominated norm and Player 2's action in the unique Bayesian Nash equilbrium.*

Proof. As before, $\hat{V}(U, P) = 0$, but in the Involuntary Ultimatum game $\hat{V}(U, A) = 10 - 6\alpha > 0$ rather than $10 - 6\alpha - 3\mu < 0$, so the norm reverses. Player 2's behavior follows from the fact that A both yields the highest material payoff to Player 2 and yields no blame. ■

Falk, Fehr and Fischbacher (2003) conduct a closely related experiment. One treatment is exactly the binary Ultimatum situation considered above; another treatment is a binary Dictator situation in which Player 1 has no choice and Player 2 chooses directly between the allocations $x_P = (0, 0)$ and $x_A = (8, 2)$. From our model's point of view, this experiment is identical to the binary version of Blount's experiment. Whether Player 1 has no choice or the choice is made by a computer does not matter for the norm or for Player 2's incentives.

[36] Blount's experiment considers a standard (non-binary) Ultimatum situation, but this is unimportant for our argument.

Like Blount (1995), Falk, Fehr and Fischbacher (2003) find that Player 1's choice set is crucial for Player 2's decision. Player 2's propensity to play P is 2.5 times greater when Player 1 can choose to play F than when Player 1 has no choice.[37]

Our model offers a rationalization of the findings of Blount (1995) and Falk, Fehr and Fischbacher (2003). By contrast, most models of other-regarding preferences imply that Player 2's propensity to play P is the same across the two treatments.[38] For example, this is a feature of Fehr and Schmidt (1999). An exception is Levine (1998), who assumes that Player 2's concern for Player 1 depends on what Player 2 believes about Player 1's type.

Recent work by Bartling and Özdemir (2017) finds that people vary greatly in their views regarding norms in binary ultimatum games.[39] In our view, this is not so much an objection to our model as a reminder that culture is not uniform.

5.4 A Trust situation

Our final example is the binary Trust situation (Figure 10), adapted from Bohnet et al (2008). Player 1 first chooses either an outside option, $s_1 = O$, which ends the situation with even payoffs (10,10), or to trust, $s_1 = T$, which continues the situation. If Player 1 trusts, Player 2 has the choice between reciprocating, $s_2 = R$, which yields payoffs (15,15), or being selfish, $s_2 = S$, which yields (8,22). The core value function yields $V(O,R) = V(O,S) = 20$, $V(T,R) = 30$ and $V(T,S) = 30 - 14\alpha$. The first result follows immediately.

Proposition 14 *In the binary Trust situation, the unique norm is* (T,R).

Consider next how players are likely to behave. Since $\alpha < 5/7$ and $\mu > (10 - 14\alpha)/5$, we have $\hat{V}(O,S) = 20$, $\hat{V}(O,R) = 20$, $\hat{V}(T,R) = 30$ and $\hat{V}(T,S) = 30 - 14\alpha - 7\mu$. Note that $V(O,S) < V(T,S)$ but $\hat{V}(O,S) > \hat{V}(T,S)$. Then, the Trust game utilities become as in Figure 11.

In order to make the problem interesting, we assume that the distribution of Player 2 types is not too extreme. More precisely, let $D(1/(2\alpha+\mu)) \in (10/(14\alpha+7\mu), 5/7)$ in what follows.[40]

[37] When Player 1 can play F, they find a rejection rate of 44.4 percent; when Player 1 has no choice, they find a rejection rate of 18 percent.

[38] Even some of the models of reciprocity that allow preferences to depend on beliefs, like Rabin (1993) and the extension by Dufwenberg and Kirchsteiger (2004), fail to account for the smaller propensity to play P in the Involuntary version.

[39] Bartling and Özdemir (2017) elicit subjects' views about the appropriateness of accepting low offers in a binary Ultimatum game. The modal response is that accepting the low offer is 'neutral: neither socially inappropriate nor appropriate' while 38 percent of the subjects rate the decision to accept the low offer as either 'very' or 'somewhat socially appropriate,' and 28 percent choose 'very' or 'somewhat socially inappropriate.'

[40] If $D(1/(2\alpha+\mu)) > 5/7$, Player 2 is so likely to play S no Player 1 type would play T. If instead $D(1/(2\alpha+\mu)) < (10/(14\alpha+7\mu)$, then all Player 1 types would play T.

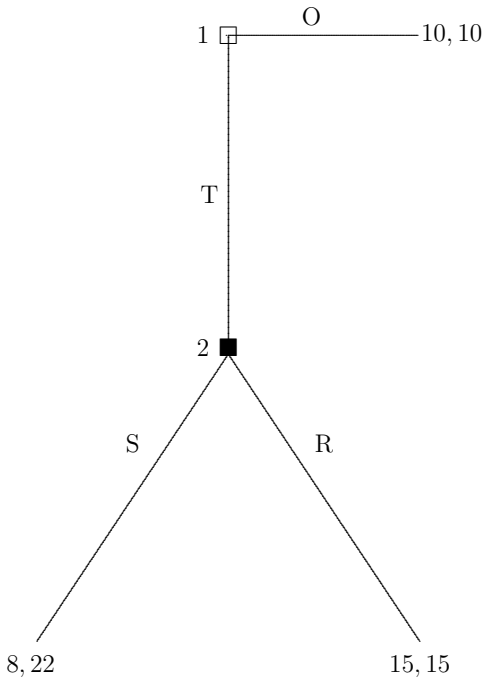

Figure 10: Trust situation

	S	R
O	$10, 10$	$10 - 10\delta_1, 10$
T	$8 - (14\alpha + 7\mu - 10)\delta_1, 22 - (14\alpha + 7\mu)\delta_2$	$15, 15$

Figure 11: Trust game

Proposition 15 *In the unique Bayesian Nash equilibrium of the Trust game, (i) Player 2 plays R if and only if*

$$\delta_2 \geq \delta^{**} = \frac{1}{2\alpha + \mu};$$

(ii) Player 1 plays T if and only if

$$\delta_1 \leq \frac{5 - 7D\left(\delta^{**}\right)}{(14\alpha + 7\mu)D\left(\delta^{**}\right) - 10}.$$

Proof. Player 2 prefers R to S if and only if

$$15 \geq 22 - (14\alpha + 7\mu)\delta_2,$$

or, equivalently,

$$\delta_2 \geq \delta^{**} = \frac{1}{2\alpha + \mu}.$$

Player 1 prefers T if and only if

$$D\left(\delta^{**}\right)\left(8-(14\alpha+7\mu-10)\delta_1\right)+\left(1-D\left(\delta^{**}\right)\right)15 \geq 10D\left(\delta^{**}\right)+\left(1-D\left(\delta^{**}\right)\right)(10-10\delta_1).$$

Simplifying and using the parameter restriction on $D(1/(2\alpha+\mu))$, the inequality becomes

$$\delta_1 \leq \frac{5-7D\left(\delta^{**}\right)}{(14\alpha+7\mu)D\left(\delta^{**}\right)-10}$$

as claimed. ■

Note that Player 1 only trusts if the own decency is not too large. The intuition is that trusting is privately profitable in expectation, but also allows a significant probability that a relatively indecent Player 2 obtains a large ill-gotten gain.

Bohnet and Zeckhauser (2004) compare Player 1's behavior the above binary trust situation to a situation in which everything is the same except that Player 2's choice between R and S is delegated to a computer. The computer's choice probabilities are set equal to the average choice probabilities in the set of subjects playing in the role of Player 2. That is, $D\left(\delta^{**}\right)$ is kept constant. Call this the Involuntary Trustworthiness situation. Although Player 2's behavior is the same in the Involuntary Trustworthiness situation as in the original Trust situation, Player 1's behavior changes; the frequency of the trusting action T goes up. Our model makes sense of this phenomenon.

Proposition 16 *Player 1 always plays T in the Involuntary Trustworthiness situation.*

Proof. Since Player 2 no longer controls the choice of S versus R, there is no ill-gotten gain for Player 2. Thus, Player 1 plays T if and only if

$$D\left(\delta^{**}\right)\left(8-(14\alpha-10)\delta_1\right)+\left(1-D\left(\delta^{**}\right)\right)15 \geq D\left(\delta^{**}\right)\left(10-\delta(10-14\alpha)\right)+\left(1-D\left(\delta^{**}\right)\right)(10-10\delta_1).$$

Some algebra simplifies the condition to

$$\delta^{**}((28\alpha-10)D(\delta^{**})-10) \leq 5-7D(\delta^{**}).$$

This condition always holds if the left-hand side is negative. A sufficient condition is that the left-hand side is negative when $D(\delta^{**}) = 5/7$ (the highest value we consider), which is equivalent to requiring $\alpha < 22/28$; this latter condition is satisfied due to our condition $\alpha < 5/7$. ■

The intuition is simple. What keeps Player 1 from trusting is only the concern that Player 2 may betray the trust and thereby create inequality and ill-gotten gains, something that a decent Player 1 dislikes. In the computer treatment, there are no ill-gotten gains, so this concern is mitigated.

Bohnet and Zeckhauser (2004) and Bohnet et al (2008) interpret such a change in behavior as an individual-level "betrayal aversion." Our model suggests that betrayal aversion could be a special case of a broader social aversion to ill-gotten gains.

The above analysis may also shed light on the observation by Glaeser et al (2000) that survey measures of trust (the extent two which people believe that others may be trusted) correlate poorly with experimental trusting behaviors but well with experimental trustworthiness. Since the survey trust measure asks about beliefs, our model does not address these correlations directly; we assume equilibrium beliefs. However, if we add the empirically grounded assumption that people have a tendency to think that others are like themselves (e.g., Iriberri and Rey-Biel, 2013), the result follows: On one hand, our model says that greater decency reduces trust because the concern for ill-gotten gains is greater, but on the other hand, greater decency increases trust because of greater optimism. Since greater decency is associated both with greater optimism and greater trustworthiness, the correlation between survey trust and experimental trustworthiness is unambiguous.

However, binary Trust experiments do produce one important regularity that our model fails to emulate. McCabe, Rigdon, and Smith (2003) compare behavior of Player 2 in the binary Trust situation described above with behavior of Player 2 when Player 1 did not have the opportunity to play O – let us call the latter case the Involuntary Trust situation.[41] They observe that Player 2's trustworthiness is lower when Player 1 did not have a choice – an instance of true positive reciprocity. By contrast, our model predicts that Player 2's behavior will be the same in the two situations, because both the distributional concerns and the ill-gotten gains are the same.

We conjecture that one reason for this predictive failure is that the model does not distinguish between different types of ill-gotten gains. In reality, Player 2 may be condemned more harshly for taking an extra payoff of seven at Player 1's expense when Player 1 "owned" (could have protected) two of these payoff units, than when Player 1 has no such protection option. Rather than a fundamental flaw of the model, this might just be an indication that entitlements also belong in the social value function.[42]

Comparing the Ultimatum and trust situations, we note that negative reciprocity (in the Ultimatum situation) is caused by individuals caring about ill-gotten gains. Effectively, the payoffs of someone who has violated a social norm are given a lower weight. In contrast, positive reciprocity (in the Trust situation) is caused by individuals caring about the core social values. The payoffs of someone who has respected a social norm, or done more than the norm requires, is not given a higher weight.

[41] For a similar experiment in a non-binary Trust setting, see Cox (2004).

[42] For a recent experiment that disentangles different potential explanations for positive reciprocity in a non-binary Trust setting, see Cox, Kerschbamer, and Neururer (2016). For experimental work on the role of entitlements in fairness judgments, see, e.g., Cappelen et al (2007) and the references therein.

6 Final Remarks

We have presented a simple model of decency. We argue that the model offers a useful framework for understanding social norms as well as for rationalizing observed duty-based moral behavior. To ascertain the empirical relevance of the framework, we have proposed a structural version of the model that applies to settings involving "manna from heaven." The model's predictions are consistent with experimental regularities that elude purely passion-based models.

A natural next step is to extend the structural model to cover other settings, and especially to introduce a role for entitlements, both from general principles (as in Cappelen et al, 2007) and from specific agreements (as in Krupka, Leider, and Jiang, 2016). The model can also be extended in many other directions. We mention three. First, individuals usually belong to many different and partly overlapping groups, whose understandings and values are not all the same. Evoking one of these several identities therefore has the potential to affect behavior. Second, values and understandings are dynamic. Within a group, the values and understandings can change due to external events as well as through negotiation or leadership. Third, individuals from one group often interact with individuals that they know belong to another group. What understandings and values guide behavior in the presence of such observed group differences? Addressing these questions will allow the model to confront findings from the large literature on social identity.

From an empirical point of view, we think the most pressing task is to jointly evaluate duty-based and passion-based motives. How large are their relative contributions to explaining observed moral behavior?

We finally hope that the model can facilitate the conversation between economists and sociologists. Duesenberry (1960, p.233) famously quipped: "Economics is about individuals' choices, sociology about how individuals don't have any choices to make." Even if there is much truth to this statement, it does not imply that sociologists and economists fundamentally differ in their beliefs regarding people's freedom of choice. It could just be that sociologists have paid more attention to social situations.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2016). Preferences for Truth-Telling, *Econometrica* forthcoming.

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

Andreoni, J. (1989). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving, *Economic Journal* 100, 464-477.

Andreoni, J. (1990). Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence, *Journal of Political Economy* 97(6), 1447-58.

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.

Arrow, K. (1974). *The Limits of Organization,* New York: W.W. Norton.

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

Bartling, B. and Fischbacher, U. (2012). Shifting the Blame: On Delegation and Responsibility, *Review of Economic Studies* 79(1), 67-87.

Bartling, B. and Özdemir, Y. (2017). The Limits to Moral Erosion in Markets: Social Norms and the Replacement Excuse, manuscript, University of Zurich.

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). Identity, Morals, and Taboos: Beliefs as Assets, *Quarterly Journal of Economics* 126(2), 805-855.

Berger, P. L. and Luckmann, T. (1966). The Social Construction of Reality: A Treatise in the Sociology of Knowledge, New York: Doubleday.

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.

Blount, S. (1995). When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences, *Organizational Behavior and Human Decision Processes* 63(2), 131-144.

Bohnet, I., Greig, F., Herrmann, B. and Zeckhauser, R. (2008). Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States, *American Economic Review* 98(1), 294-310.

Bohnet, I. and Zeckhauser, R, (2004). Trust, Risk and Betrayal, *Journal of Economic Behavior and Organization* 55(4), 467-484.

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

Breitmoser, Y. and Vorjohann, P. (2017). Welfare-based Altruism, manuscript, Humboldt University Berlin.

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.

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.

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.

Coleman, J.S. (1988). Social Capital in the Creation of Human Capital, *American Journal of Sociology* 94 (Supplement), S95-S120.

Coleman, J.S. (1990). *Foundations of Social Theory*, Cambridge MA: The Belknap Press of Harvard University Press.

Conrads, J. and Irlenbusch, B. (2013). Strategic Ignorance in Ultimatum Bargaining, *Journal of Economic Behavior and Organization* 92, 104-115.

Cox, J.C. (2004). How to Identify Trust and Reciprocity? *Games and Economic Behavior* 46, 260-281.

Cox, J. C., Friedman, D., & Gjerstad, S. (2007). A Tractable Model of Reciprocity and Fairness. *Games and Economic Behavior* 59(1), 17-45.

Cox, J.C., Kerschbamer, R., and Neururer, D. (2016). What Is Trustworthiness and What Drives It? *Games and Economic Behavior* 98, 197-218.

Crockett, M. J., Clark, L., Lieberman, M. D., Tabibnia, G., and Robbins, T. W. (2010). Impulsive Choice and Altruistic Punishment are Correlated and Increase in Tandem with Serotonin Depletion, *Emotion* 10(6), 855-862.

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.

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.

Dufwenberg, M., Heidhues, P., Kirchsteiger, G., Riedel, F. and Sobel, J. (2011). Other-Regarding Preferences in General Equilibrium, *Review of Economic Studies* 78(2), 613-639.

Dufwenberg, M. and Kirchsteiger, G. (2004). A Theory of Sequential Reciprocity, *Games and Economic Behavior* 47(2), 268-298.

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

Ellingsen, T., Johannesson, M., Mollerstrom, J., and Munkhammar, S. (2011). Social Framing Effects: Preferences or Beliefs? *Games and Economic Behavior* 76(1), 117-130.

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

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

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the Nature of Fair Behavior, *Economic Inquiry* 41(1), 20-26.

Falk, A., Fehr, E., Fischbacher, U. (2008). Testing Theories of Fairness: Intentions Matter, *Games and Economic Behavior* 62(1), 287-303.

Falk, A. et al (2018). Global Evidence on Economic Preferences, *Quarterly Journal of Economics* 133(4), 1645-1692.

Fehr, E. and Schurtenberger, I. (2018). Normative Foundations of Human Cooperation, *Nature Human Behavior* forthcoming.

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.

Foster, D.P. and Young, H.P. Stochastic Evolutionary Game Dynamics, *Theoretical Population Biology* 38, 219-232.

Freddi, E. (2017). Do People Avoid Morally Relevant Information? Evidence from the Refugee Crisis, *Job Market Paper* Stockholm School of Economics.

Friedman, M. (1970). The Social Responsibility of Business Is to Increase Its Profits, *New York Times Magazine* 13 September 1970.

Gabaix, X. (2014). A Sparsity-based Model of Bounded Rationality, *Quarterly Journal of Economics*129(4), 1661-1710.

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

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

Glaeser, E., Laibson, D.I., Scheinkman, J.A. and Soutter, C.L. (2000). Measuring Trust, *Quarterly Journal of Economics* 115(3), 811-846.

Glazer, A., Konrad, K.A. (1996). A Signaling Explanation of Charity, *American Economic Review 86, 1019-1028.*

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

Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information Avoidance, *Journal of Economic Literature* 55(1), 96-135.

Gospic, K., Mohlin, E., Fransson, P., Petrovic, P., Johannesson, M., & Ingvar, M. (2011). Limbic Justice—Amygdala Involvement in Immediate Rejection in the Ultimatum Game. *PLoS Biology*, 9(5).

Gouge, W. (1622). *Of Domesticall Duties: Eight Treatises* London: John Haviland, for William Bladen.

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

Haidt, J. (2003). The Moral Emotions, Chapter 45 in R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (eds.), *Handbook of Affective Sciences* 11, 852-870, Oxford: Oxford University Press.

Harrod, R.F. (1936). Utilitarianism Revised, *Mind* 45(178), 137-156.

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

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.

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

Iriberri, N. and Rey-Biel, P. (2013). Elicited Beliefs and Social Information in Modified Dictator Games: What Do Dictators Believe Other Dictators Do? *Quantitative Economics* 4, 515-547.

Jehiel, P. (2005). Analogy-based Expectation Equilibrium, *Journal of Economic Theory,* 123(2), 81-104.

Kahn, V. (1999). "The Duty to Love": Passion and Obligation in Early Modern Political Theory, *Representations* 68(Autumn), 84-107.

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

Kandori, M., Mailath, G,J., and Rob, R. (1993). Learning, Mutation, and Long Run Equilibria in Games, *Econometrica* 61(1), 29-56.

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

Knutsson, M., Martinsson, P., Wollbrant, C. (2013). Do People Avoid Opportunities to Donate? A Natural Field Experiment on Recycling and Charitable Giving. *Journal of Economic Behavior and Organization* 93, 71-77.

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

Krupka, E.L., Leider, S., and Jiang, M. (2016). A Meeting of the Minds: Informal Agreements and Social Norms. *Management Science* 63(6), 1708-1729.

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.

Kuzmics, C. (2011). On the Elimination of Dominated Strategies in Stochastic Models of Evolution with Large Populations, *Games and Economic Behavior* 72(2), 452-466.

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

Levine, D. K. (1998). Modelling Altruism and Spitefulness in Experiments, *Review of Economic Dynamics* 1, 593-622.

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. (2009). Social Preferences: Some Thoughts from the Field. *Annual Review of Economics* 1, 563-579.

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

Mailath, G., Morris, S., and Postlewaite, A. (2017). Laws and Authority, *Research in Economics* 71 (1), 32-42.

Malmendier, U., and Schmidt, K.M. (2017). You Owe Me. *American Economic Review*, 107 (2), 493-526.

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

Mansbridge, J. (1998). Starting With Nothing: On the Impossibility of Grounding Norms Solely in Self-Interest, Chapter 5 in A. Ben-Ner and L. Putterman (eds.) *Economics, Values, and Organization,* Cambridge: Cambridge University Press.

March, J. (1994). *A Primer on Decision-Making: How Decisions Happen,* New York: Free Press.

March, J. and Olsen, J.P. (1989). *Rediscovering Institutions,* New York: Free Press.

March, J. and Olsen, J.P. (2011). The Logic of Appropriateness, in R.E. Goodin (ed.) *The Oxford Handbook of Political Science,* Oxford: Oxford University Press.

McCrae, R.R. and Costa, P.T. (2003). *Personality in Adulthood: A Five-Factor Perspective* 2nd Edition, New York: Guildford Press.

Mohlin, E. (2014) "Optimal Categorization", *Journal of Economic Theory,* 152, pp. 356–381.

Myerson, R.B. (1991). *Game Theory: Analysis of Conflict,* Cambridge MA: Harvard University Press.

Nachbar, J., (1990). 'Evolutionary' Selection Dynamics in Games: Convergence and Limit Properties. *International Journal of Game Theory* 19, 59-89.

Opp, K.D. (1982). The Evolutionary Emergence of Norms. *British Journal of Social Psychology* 21(2), 139-149.

Östling, R., Wang, J.T., Chou, E.Y. and Camerer, C.F. (2011). Testing Game Theory in the Field: Swedish LUPI Lottery Games, *American Economic Journal: Microeconomics* 3 (3), 1-33.

Parsons, T. (1951). *The Social System,* New York: The Free Press.

Pelto, P.J. (1968). The Difference Between "Tight" and "Loose" Societies, *Trans-action* 5(5), 37-40.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics, *American Economic Review* 83(5), 1281–1302.

Rabin M (1994). Cognitive dissonance and social change, *Journal of Economic Behavior and Organization,* 23, 177-194.

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

Ross, L., and Nisbett, R.E. (1991). *The Person and the Situation: Perspectives of Social Psychology,* 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.

Samuelson, L. (1994), Stochastic Stability in Games with Alternative Best Replies, *Journal of Economic Theory* 64(1), 35-65.

Sandholm, W.H. (2010). *Population Games and Evolutionary Dynamics,* Cambridge MA: MIT Press.

Segal, U. and Sobel, J. (2007). Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings, *Journal of Economic Theory* 136(1), 197-216.

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

Spiekermann, K., and Weiss, A. (2016) Objective and Subjective Compliance: A Norm-Based Explanation of 'Moral Wiggle Room'. *Games and Economic Behavior* 96: 170-183.

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

Tajfel, H. (1982). Social Psychology of Intergroup Relations, *Annual Review of Psychology* 33, 1-39.

Thibaut, J.W., and Kelley, H.H. (1959). *The Social Psychology of Groups.* Oxford: John Wiley.

Thomas, W.I. and Thomas, D.S. (1928) *The Child in America: Behavior Problems and Programs,* New York: Knopf.

Ullman-Margalit, E. (1977). *The Emergence of Norms,* Oxford: Clarendon Press.

Van der Weele, J. (2014). Inconvenient Truths: Determinants of Strategic Ignorance in Moral Dilemmas, Manuscript, University of Amsterdam.

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

Williamson, O.E. (1975). *Markets and Hierarchies,* New York: Free Press.

Young, H.P. (2015). The Evolution of Social Norms, *Annual Review of Economics* 7, 359-387.

7 Appendix: Non-linear disutility of blame

Let the cost of blame be

$$U^b\left(b_i\left(s\right)\right) = \frac{\left(b_i\left(s_i, s_{-i}\right)\right)^2}{\max_{\tilde{s}_i} b_i\left(\tilde{s}_i, s_{-i}\right)} + \phi \mathbf{1}_{\{b_i(s_i,s_{-i})>0\}} \cdot \max_{\tilde{s}_i} b_i\left(\tilde{s}_i, s_{-i}\right).$$

This cost-function has both a variable and a fixed component. The variable component is continuously increasing and convex in blame. The expression $\max_{s_i} b_i\left(s_i, s_{-i}\right)$ in the denominator corresponds to the maximal blame that the player could effect, and serves as a normalization. The fixed cost is incurred for any positive blame, and is equal to $\max_{\tilde{s}_i} b_i\left(\tilde{s}_i, s_{-i}\right)$. The parameter ϕ measures the importance of the fixed component relative to the variable component. We assume $\phi < 1$. (The case $\phi = 1/2$ yields attractive solutions.)

We maintain the assumptions that material utility is linear, i.e. $U^z\left(x\left(s\right)\right) = x_i\left(s\right)$, and total utility is additively separable, i.e.

$$U_i\left(s\right) = x\left(s\right) - \delta_i U^b(b_i\left(s\right)),$$

where blame (and social value) is defined as before.

7.1 Dictator situations

Consider the Dictator situation, as described in the main text. The unique value-maximizing allocation, and thus the only social norm, is the equal split, $s_1 = 5$. Blame is

$$\begin{aligned} b_1\left(s_1\right) &= \max_{\bar{s}_1} V\left(x\left(\bar{s}_1\right)\right) - V\left(x\left(s_1\right)\right) \\ &= 10 - \left(10 - \alpha\left|s_1 - \left(10 - s_1\right)\right|\right) \\ &= \alpha\left|2s_1 - 10\right|, \end{aligned}$$

and maximal blame is $\max_{\tilde{s}_1} b_1\left(\tilde{s}_1\right) = 10\alpha$, so that

$$U^b\left(b_1\left(s\right)\right) = \frac{\left(\alpha\left|2s_1 - 10\right|\right)^2}{10\alpha} + \phi \mathbf{1}_{\{s_i \neq 5\}} \cdot 10\alpha.$$

The dictator maximizes $U_1\left(s_1\right) = s_1 - \delta_1 U^b\left(b_1\left(s\right)\right)$. For $s_1 < 5$, utility is increasing in s_1, and for $s_1 > 5$,

$$U_1\left(s_1\right) = s_1 - \delta_1\left(\frac{\alpha\left(2s_1 - 10\right)^2}{10} + \phi 10\alpha\right).$$

Inspecting the latter expression yields the result:

Proposition 17 *In the Dictator situation the amount kept is*

$$s_1 = \begin{cases} 10 & \text{if } \delta_1 \leq \frac{1}{4\alpha}; \\ 5 + \frac{5}{4\alpha\delta_1} & \text{if } \frac{1}{4\alpha} < \delta_1 < \frac{1}{4\alpha\sqrt{\phi}}; \\ 5 & \text{if } \delta_1 \geq \frac{1}{4\alpha\sqrt{\phi}}. \end{cases}$$

Proof. For $s_1 > 5$, we have

$$U_1'(s_1) = 1 - \delta_1 \frac{4\alpha(2s_1 - 10)}{10},$$

and $U_1''(s_1) < 0$. Thus if the dictator finds it optimal to set $s_1 \in (5, 10)$ then the optimal s_1 solves

$$1 = \delta_1 \frac{4\alpha(2s_1 - 10)}{10},$$

or equivalently $s_1 = 5 + \frac{5}{4\alpha\delta_1}$. Note that this larger than 5 for any finite $\alpha\delta_1$, and less than 10 for any $\delta_1 > \frac{1}{4\alpha}$. Conversely, if $\delta_1 \geq \frac{1}{4\alpha}$ then $U_1\left(5 + \frac{5}{4\alpha\delta_1}\right) \leq U_1(10)$ with equality only at $\delta_1 = \frac{1}{4\alpha}$. Utility of $s_1 = 5 + \frac{5}{4\alpha\delta_1}$ is

$$\begin{aligned} U_1\left(5 + \frac{5}{4\alpha\delta_1}\right) &= 5 + \frac{5}{4\alpha\delta_1} - \delta_1 \left(\frac{\alpha\left(2\left(5 + \frac{5}{4\alpha\delta_1}\right) - 10\right)^2}{10} + \phi 10\alpha \right) \\ &= 5 + \frac{5}{4\alpha\delta_1} - \delta_1 \left(\frac{5}{8\alpha\delta_1^2} + \phi 10\alpha \right). \end{aligned}$$

In comparison the utility from $s_1 = 5$ is $U_1(5) = 5$. Hence $U_1\left(5 + \frac{5}{4\alpha\delta_1}\right) > U_1(5)$ iff

$$\frac{5}{4\alpha\delta_1} > \delta_1 \left(\frac{5}{8\alpha\delta_1^2} + \phi 10\alpha \right),$$

or equivalently

$$\delta_1 < \frac{1}{4\alpha\sqrt{\phi}}.$$

Thus if $\frac{1}{4\alpha\sqrt{\phi}} \leq \delta_1$ it is optimal for Player 1 to set $s_1 = 5$, whereas if $\frac{1}{4\alpha} < \delta_1 < \frac{1}{4\alpha\sqrt{\phi}}$ then it is optimal to set $s_1 = 5 + \frac{5}{4\alpha\delta_1} \in (5, 10)$.

The utility from $s_1 = 10$ is $U_1(10) = 10 - \delta_1(1 + \phi)10\alpha$, so that $U_1(10) > U_1(5)$ iff $5 > \delta_1(1 + \phi)10\alpha$ or equivalently $\delta_1 < \frac{1}{(1+\phi)2\alpha}$. Note that $\delta_1 \leq \frac{1}{4\alpha}$ implies $\delta_1 < \frac{1}{(1+\phi)2\alpha}$ by the assumption that $\phi < 1$. Thus if $\delta_1 \leq \frac{1}{4\alpha}$ then it is optimal to set $s_1 = 10$. ■

Next, consider the Dictator situation with an exit option. First suppose the exit choice is seen as a choice within a social situation. We obtain the same result as in the main text: Everyone sticks to their original choice, at least under the assumption that $\alpha < 1$, which implies that the exit option creates a lower social value than choosing $s_1 = 9$ in the standard

dictator situation.

Proposition 18 *Suppose the exit choice is seen as a choice within a social situation. Suppose $\alpha < 1$. Then for all values of δ_1 the original choice s_1 is preferred.*

Proof. Available upon request. ■

Suppose the exit choice is seen as a choice between social situations. The result is similar to the result in the main text: those with low enough δ maintain their choice $s_1 = 10$ whereas everyone else chooses to exit.

Proposition 19 *Suppose the exit choice is seen as a choice between social situations, not within. Then,*

$$\tilde{s}_1 = \begin{cases} s_1 = 10 & \text{if } \delta_1 \text{ and } \delta_1 < 1/10\alpha\,(1+\phi) \\ e & \text{otherwise.} \end{cases}$$

Proof. Available upon request. ■