Ellingsen, Tore; Mohlin, Erik

Working Paper

Decency

Working Paper, No. 2019:3

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

Suggested Citation: Ellingsen, Tore; Mohlin, Erik (2019) : Decency, Working Paper, No. 2019:3, Lund University, School of Economics and Management, Department of Economics, Lund

This Version is available at:
https://hdl.handle.net/10419/260273

Working Paper 2019:3

Department of Economics
School of Economics and Management

Decency

Tore Ellingsen
Erik Mohlin

February 2019

LUND
UNIVERSITY

Decency*

Tore Ellingsen† Erik Mohlin‡

February 21, 2019

Abstract

We develop a formal theory of decency. Shared values and understandings give rise to social norms. Norms may mandate collectively optimal behavior, but 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, Pol Campos-Mercade, Andrew Caplin, Ernst Fehr, Sebastian Fehrler, Erik Gaard Kristiansen, Robert Östling, Zoltán Rácz, Alexandros Rigos, 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, passion and decency 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).

In this paper, our main purpose is to construct a simple and portable formal model of decency. As an indication of the model's explanatory power, we demonstrate that it offers a unified account for a variety of behavioral regularities that defy standard passion-based models. In addition to reluctant charity, that the model was designed to accommodate, the model explains willful ignorance of externalities (e.g., Dana, Weber, and Kuang, 2007; Freddi, 2019), intention-based negative reciprocity (e.g., Blount, 1995; Falk, Fehr and Fischbacher, 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 en-

[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.

trances 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.[2] In comparison with innate moral passions, decency is thus more immediately tied to cultural variation in moral behavior.[3] 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.[4]

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 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

[2]For references and a recent well-identified study of this process, see Kosse et al (2019).

[3]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).

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

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).[5] 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.[6] By making explicit the distinction between social and non-social situations, the model highlights an obstacle to generalizing from laboratory experiments to field settings. The generalization requires that the sociality of the situation is preserved. Thus, the model immediately implies that lab-to-field generalization is harder to accomplish when social values and understandings matter for individual decision-making (as in Levitt and List, 2007; List, 2009; Galizzi and Navarro-Martinez, in press), than when social values are unimportant (as in Östling et al, 2011, and the references therein).

Related literature. 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 Five personality traits, altruism is a facet of Agreeableness whereas dutifulness is a facet of Conscientiousness; see, e.g., McCrae and Costa (2003). In fact, a large literature suggests that there is a sixth dimension of personality that might well be called decency. It usually goes under the name of honesty-humility, but has also been called morality or selfishness (when scaled inversely); see Diebels, Leary, and Chon (2018). One objection to including this sixth dimension in the personality inventory has been that is does not seem to be defined entirely identically across countries. Our model offers a possible solution to this quandary by suggesting that the personality trait itself is stable, but that the social values to which the individuals adapt differ somewhat between countries.[7]

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

[5]The relationship between objective social reality and subjective understanding of reality is an age-old topic in philosophy. In more recent 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). Here, we take for granted that social reality has been created, leaving aside the questions of how and why.

[6]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.

[7]Becker et al (2012) observe that there is only a low correlation between Big Five personality traits and behavior in typical behavioral economics experiments. It might be worthwhile redoing the analysis with the honesty-humility trait included.

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

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 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

[9] For a brief history of cultural tightness concepts, see Pelto (1968). An influential recent 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.

[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).

identity.[12]

Compared to these previous approaches to internalized social norms, we establish a more basic framework in which the norms themselves derive from general 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 the same preferences apply to all situations. In particular, like Dillenberger and Sadowski (2012), we explicitly allow for the possibility that decision-makers can seek or avoid situations in which social 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 likewise utilizes a refinement of Nash equilibrium, but does not consider the issue of evolutionary selection.[16]

The paper is organized as follows. Section 2 presents the formal concepts. An important definition is the notion of an injunctive social norm, which we take to be a profile of actions such that each individual maximizes social value conditional on the actions taken by others. Another important definition is the notion of blameworthiness, which is measured as the loss of social value that an individual causes. The section also provides our simplest structural model of social values, based on desire for efficiency and equality only. Section 3 applies the model to explain evidence from Dictator experiments with and without exit options and from Moral Wiggle Room experiments. Section 4 extends the model to incorporate a dislike for ill-gotten gains; this extension proves crucial for explaining both negative reciprocity and lack of trust, as becomes clear when we apply the model to various Ultimatum and Trust experiments. Section 5 concludes.

[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.

[16]The modern literature on evolutionary game theory and morality has several different strands; see for example Alger and Weibull (2013) and Binmore (2005).

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.

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".[17]

Within a culture, some situations are considered to be social, and other situations are non-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 in a social situation, the individual is blameworthy and will suffer some guilt. Conversely, if the situation is considered non-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. While we shall mostly take for granted that individuals understand whether a situation is social or not, we recognize that this is assumption is not always satisfied in practice. In unfamiliar situations, such as interactions with people from other cultures or in neutrally framed laboratory experiments 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 is a set of n players, $S = \times_i S_i$ is a 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.[18] Let $\Sigma = \times_i \Sigma_i$ denote the set of mixed strategy profiles, with σ being a typical element.

Games. The standard definition of a game with complete information assumes that

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

[18] 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.

Player i's preferences are captured by a von Neumann-Morgenstern utility function, $U_i : Z \to \mathbb{R}$, with $u_i(\sigma) := E_\sigma [U_i (x (s))]$. Here, we shall allow richer (more sociological) preferences, that depend on the player's blameworthiness, b_i (an endogenous quantity to be defined in Section 2.4), while retaining the key property of expected utility theory that $u_i(\sigma) := E_\sigma [U_i (x (s) , b_i)]$. As usual, we let $u_i(s)$ denote the utility associated with the pure strategy profile s. A complete information game is a tuple $G = \langle F, u \rangle$.

In order to capture heterogeneous morality, and players' associated uncertainty about others' morality, we also consider a restricted class of games with incomplete information, where players' preferences depend on their own type, but not directly on the type of their opponents. Letting $T = \times_i T_i$ denote the set of type profiles, Player i's utility function is then $U_i : Z \times T_i \to \mathbb{R}$, with $u_i(\sigma, t_i) := E_\sigma [U_i (x (s) , b_i, t_i)]$. Let players share the same prior beliefs $p(t)$ about the distribution of types. An incomplete information game (or a Bayesian game) is thus a tuple $G^B = \langle F, T, p, u \rangle$.

Solution concepts. For games with complete information, 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 $i \in N$,*

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

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}.*

For games with incomplete information (with common priors and no utility-dependence on opponent's types), we accordingly consider Bayesian Nash equilibrium.

Definition 3 *A strategy profile $\sigma^*(t)$ is a Bayesian Nash equilibrium of a game G^B if, for all $t \in T$ and $i \in N$,*

$$\sigma_i^*(t_i) \in \arg\max_{\sigma_i \in \Sigma_i} E_{t_{-i}}[u_i(\sigma_i(t), \sigma_{-i}^*(t_{-i}))].$$

If players move sequentially, the actions of early movers may provide clues about their type. This would be important, and require equilibrium refinement, if one player's utility were to depend directly on another player's type. But since we abstract from such dependence here, sequential moves merely simplifies the analysis by eliminating the uncertainty of later players about what earlier players do.

The solution concepts are used for generating predictions about behavior.[19] Behavior σ

[19]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, Bernergård and Mohlin (2019) and the references therein.

and situations F are easier to observe empirically than preferences u. Our objective is to make inferences about a stable but hidden function u by varying F. Therefore, it is natural to express the model's predictions in terms of σ and F, while keeping u in the background.

Definition 4 *Suppose players have complete information. A strategy profile σ is a potential convention of the situation F if it is an undominated pure strategy Nash equilibrium of the game $G = \langle F, u \rangle$.*

When the game does not have complete information (for example due to private information about preferences), we extend this definition by considering Bayesian equilibrium instead.

Definition 5 *Suppose players have incomplete information. A strategy profile σ is a potential convention of the situation F if it is a Bayesian Nash equilibrium of $G = \langle F, T, p, u \rangle$.*

Note that we use the word "convention" in the broad sense of what a player of a particular type will be doing rather than the more restrictive sense of what most players will be doing.

Remark. A common cause of confusion in the literature is that situations and games are mixed up. For example, the literature talks about experimental evidence on Dictator, Trust, and Ultimatum *games,* as if the monetary payoffs in these experimental situations corresponded to von Neumann-Morgenstern utilities.[20] We propose to reserve those names for the cases in which payoffs are considered to be utilities. When describing the monetary payoffs, we refer to Dictator, Trust, and Ultimatum *situations.*

Choice between situations. Since a situation is typically only a small excerpt of reality, players will frequently be choosing between situations. We make the assumption that such choices themselves are not a source of utility. More generally, utilities from separate games are additive.

Consider for example the a customer choosing between supermarket entrances in the study of Andreoni, Trachtman and Rao (2017). Suppose the interaction with a Salvation Army officer is considered a situation that occurs only if the customer is close to the officer. Then, the door choice itself is another situation. Assume that distance walked is a relevant concern. The additivity assumption then says that the customer's utility is the sum of the utilities from walking distance and from the donation choice within the situation with the Salvation Army officer (if the customer chooses the door where the officer stands).

For the most part, we take for granted that all players partition reality in the same way; they have perfectly shared understandings of how situations are bracketed. In future work, we hope to consider both heterogeneity and malleability of understandings.

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.

[20] If we knew that players were selfish materialists, this practice would be all right. For example, rejecting an ultimatum offer might then justly be described as a failure to play the subgame perfect equilibrium. But if we do not know players' preferences, it seems more reasonable to interpret the rejection as a deliberate act of punishment.

2.3 Social values and social norms

Social values order the set of outcomes Z in social situations. Thus, social values are expressed by a function $V : Z \to \mathbb{R}$.[21] Let the social value associated with a strategy profile σ be denoted[22]

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

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 6 *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 given what (they believe that) others will do.[23] Note the analogy with Nash equilibrium. Let $\Sigma^{PN}(F, v)$ denote the set of social norms in social situation F.

From now on, we refer simply to norms rather than social norms. Our next definition singles out the norms that maximize social value.

Definition 7 *An ideal norm is a strategy profile*

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

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.

[21] Individuals may or may not think that the values V are justified; that distinction does not matter for the positive analysis of this paper, but is important for normative analysis. In the case that people approve of V, the values could then represent the preferences of Adam Smith's impartial spectator; we are grateful to Andrew Caplin for this observation.

[22] 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.

[23] The literature sometimes adds the qualifiers *prescriptive* or *injunctive* in order to separate the norms that people ought to follow from their typical behavior – the *descriptive* norms. Here, we have already reserved the label convention for typical behavior, and so the qualifier is superfluous.

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 8 *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. See Appendix. ■

In other words, there is always an undominated norm that tells each player exactly what to do, and that norm maximizes social value.[24]

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.[25]

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

Players that fail to maximize social value in social situations are *blameworthy*. Let the blameworthiness $b_i : S \to \mathbb{R}$ of Player i equal the social loss that Player i causes,

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

According to this definition, blameworthiness depends on what others are doing. For example, in a weak-link situation (such as Stag Hunt) no player is blameworthy for taking the lowest action if at least one other player does so – but a player is blameworthy for being the only player not to take the highest action.[26]

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

[24] 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.

[25] 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.

[26] 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

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

$U_i^z : Z \to \mathbb{R}$. Observe that material utility is allowed to depend on the whole profile of material payoffs $x(s)$, not only the individual's own payoff $x_i(s)$. 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

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

Define *guilt* as the disutility of blameworthiness, i.e., $\delta_i U^b(b_i(s))$. We refer to $\delta_i \geq 0$ as Player i's degree of *decency*. Decency is the only source of preference heterogeneity that we consider, as we let $U_i^z(x(s)) = U_i^z(x_i(s))$ from now on.[27]

Suppose players take blame into account even when nobody can observe their behavior. That is, players feel guilt when they are blameworthy, not only when they are actually blamed by others. For example, guilt from blameworthiness may keep people from stealing in situations where they know that the crime could not be discovered.[28] Finally, assume that Player i maximizes the expectation of U_i, that is,[29]

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

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 feel guilty about not maximizing social value.[30] 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. In particular, there are many situations in which the lost social value associated with an individual's deviation from a norm will be greater when others do not similarly deviate. Take the example of littering. If the street is already littered, a little extra litter is hardly noticeable. If the street is completely clean, a piece of litter matters much more; in terms of the model, the own littering will have a

[27] In terms of the general model, the type t_i thus corresponds to δ_i and the set T of feasible types is a subset of $\mathbb{R}_+$.

[28] 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.

[29] 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.

[30] This is not to say that there is never specific disutility from norm-breaking in reality. A natural extension of the model is to allow norms to be highlighted through recommendations and laws, which in turn create additional blame in the form of a fixed utility cost of norm-violation. This extension would constitute a formalization of ideas about the expressive function of law (e.g., Sunstein,1996; Kahan, 1997; Cooter, 1998.)

greater impact on V.[31] Thus, it is *as if* the individual feels more compelled to comply with the norm if others also comply.[32]

Note finally that people may be blamed even if they personally comply with an ideal norm. The reason is that the own norm compliance may create social harm when others do not comply with the ideal norm. An example is driving on a motorway. The ideal social norm might instruct everyone to comply with the stated speed limit. Yet, if most motorists are driving significantly faster, a driver might well cause trouble by obeying the ideal norm rather than the established convention (which itself, despite the inefficiency, might be an undominated norm). In this case, the model says that the least blameworthy behavior is to adjust the speed to that of the surrounding traffic.

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

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

where $\alpha > 0$ and

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

That is, society puts positive value on efficiency (x^{+}) and negative value on inequality (x^{-}).

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,

[31] According to the experiments of Cialdini, Reno, and Kallgren (1990), people may be even less prone to littering when they see a single noticeable piece of litter than when they see none, because in the latter case they are not reminded of the norm against littering.

[32] 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 relevant 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 think 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.

2016; Malmendier, Velde, and Weber, 2014).

In Section 4, we enrich the model by adding to the social value function a dislike for ill-gotten gains. It is straightforward to show that all qualitative results in Section 3 hold under that richer specification, so the simplicity of the initial analysis comes at no cost.

3 Dictator situations

The Dictator situation involves two players, Player 1 and Player 2, with strategy sets $S_1 = [0, 10], S_2 = \{\emptyset\}$. 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 1 *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.

3.1 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.[33]

Suppose Player 1 views the whole Dictator experiment with exit option as a single social situation.[34] After the exit option is presented, she faces the choice set $\tilde{S}_1 = \{e, s_1\}$, where

[33]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).

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

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

Proof. See Appendix. ■

So, why do subjects exit? The model offers a straightforward resolution. The original Dictator situation 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.[35]

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 3 *Suppose the exit choice is seen as a choice between a social situation and a non-social situation. Then,*

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

Proof. See Appendix. ■

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.

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.

[35] 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.

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.[36]

3.2 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

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

[36] 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.

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 4 *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* $\delta_1 < 1/(3+10\alpha p)$.

Proof. See Appendix. ■

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 5 *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. See Appendix. ■

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 5 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.[37]

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

[37] 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.

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.

4 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 opponents' 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.[38]

[38] 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.

4.1 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

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

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

$$U_i(s) = U_i^z(x(s)) - \delta_i U^{\hat{b}}(\hat{b}_i(s)).$$

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

4.2 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.

4.3 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).

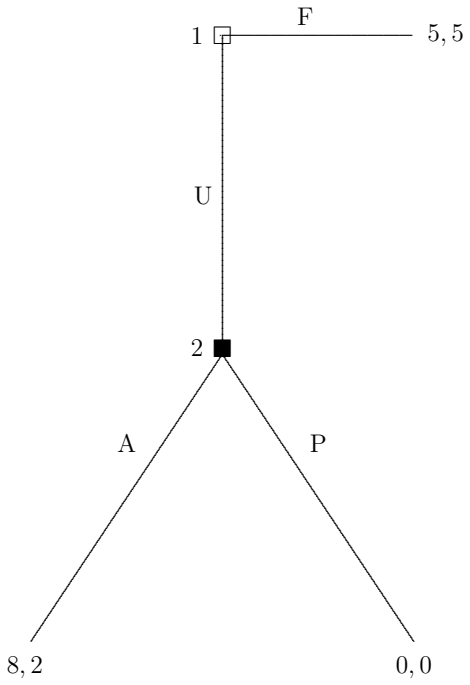

Figure 8: Ultimatum situation

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

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.

	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 situation preferences

Proposition 6 *In the Ultimatum situation, the unique norm is* (F,P).

Having derived the extended value function, the players' utilities are as in Figure 9. Let us now characterize behavior in the Ultimatum situation.

Proposition 7 *The unique potential convention of the Ultimatum situation is: (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*

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

Proof. See Appendix. ■

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.[39] 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.

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

Proposition 8 *In the Involuntary Ultimatum situation, action A by Player 2 is both the unique undominated norm and Player 2's action in the unique potential convention.*

Proof. See Appendix. ■

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. 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.[40]

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.[41] 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 situations.[42] In our view, this is not so much an objection to our approach as a reminder that culture is not uniform. Sometimes, we are uncertain not only about others' decency, but also about their values.

4.4 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 9 *In the binary Trust situation, the unique norm is* (T, R).

[40] 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.

[41] 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.

[42] Bartling and Özdemir (2017) elicit subjects' views about the appropriateness of accepting low offers in a binary Ultimatum situation. 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.'

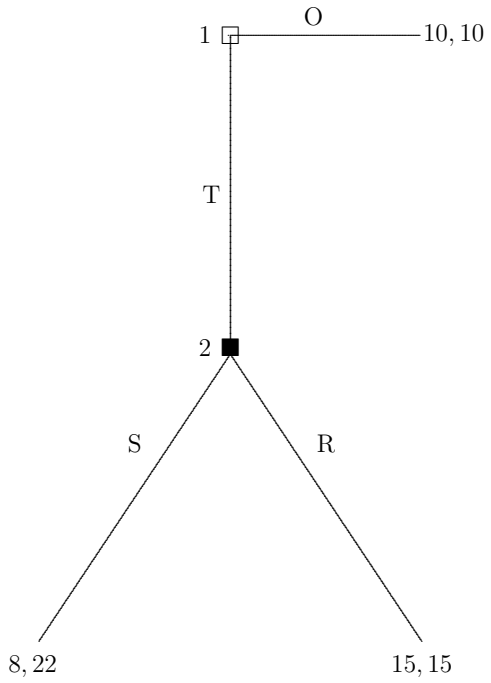

Figure 10: Trust situation

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 situation utilities become as in Figure 11.

	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 situation preferences

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.[43]

Proposition 10 *The unique potential convention of the Trust situation is (i) Player 2 plays R if and only if*

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

[43]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.

(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. See Appendix. ■

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

Proof. See Appendix. ■

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 to 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.[44] 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.[45]

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.

5 Final Remarks

Commenting on Gary Becker's individualistic analysis of fertility choice, a topic which at the time was considered outside the realm of economics, James Duesenberry (1960, p.233) famously quipped: "Economics is about individuals' choices, sociology about how individuals don't have any choices to make." He went on to explain why both perspectives are valuable. Our model of decency offers an integrated framework allowing us to analyze both perspectives simultaneously rather than separately.

To ascertain the empirical relevance of the framework, we proposed a structural version of the model that applies to settings involving "manna from heaven." The structural model's predictions are consistent with experimental regularities that elude purely passion-based models.

A natural next step is to extend the model to cover settings in which people hold entitlements, either from prior principles (as in Cappelen et al, 2007) or arising from their own agreements (as in Krupka, Leider, and Jiang, 2016). A broader challenge is to understand what drives the scope and sociality of a situation. Ideally, we seek a formal model

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

[45] 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.

of endogenous social bracketing that simultaneously explains which individuals have social responsibility and what actions these individuals consider to be part of a situation. While role theory already offers a rich set of answers to this question, the answers often appear contradictory, presumably due to the lack of formalization (Biddle, 1986).

References

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

Alger, I. and Weibull, J. (2013). Homo moralis – preference evolution under incomplete information and assortative matching, *Econometrica* 81, 2269-2302.

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, A., Deckers, T., Dohmen, T., Falk, A., and Kosse, F. (2012). The Relationship Between Economic Preferences and Psychological Personality Measures, *Annual Review of Economics* 4, 453-478.

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.

Bernergård, A. and Mohlin, E. (2019). Evolutionary Selection against Iteratively Weakly Dominated Strategies, *Games and Economic Behavior* forthcoming.

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

Biddle,B.J. (1986). Recent Developments in Role Theory, Annual Reviews of Sociology 12, 67-92.

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.

Cialdini, R.B., Reno, R.R:, and Kallgren, C.A. (1990) A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places, *Journal of Personality and Social Psychology* 58, 1015-1029.

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.

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

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

Cox, J. C., Friedman, D., and 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.

Diebels, K.J., Leary, M.R., and Chon, D. (2018). Individual Differences in Selfishness as a Major Dimension of Personality: A Reinterpretation of the Sixth Personality Factor, *Review of General Psychology* 22, 367-376.

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

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

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

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. (2019). Do People Avoid Morally Relevant Information? Evidence from the Refugee Crisis, *Review of Economics and Statistics* forthcoming.

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. and 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., and 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.

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

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., and 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.

Kosse, F., Deckers, T., Pinger, P., Schildberg-Horisch, H., and Falk, A. (2019). The Formation of Prosociality: Causal Evidence on the Role of Social Environment, *Journal of Political Economy* forthcoming.

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.

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.

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, M. (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.

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.

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

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.

Appendix A: Proofs

Proof of Theorem 2

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, $v\left(\hat{s}_i, \hat{s}_{-i}\right) = v\left(\sigma_i', \hat{s}_{-i}\right)$, meaning that $\left(\sigma_i', \hat{s}_{-i}\right)$ is also an ideal norm. Let $C\left(\sigma_i'\right)$ denote the support of σ_i'. Since $v\left(\sigma_i', \hat{s}_{-i}\right) = \sum_{s_i \in C\left(\sigma_i'\right)} \sigma_i'\left(s_i\right) v\left(s_i, \hat{s}_{-i}\right)$, the fact that v is maximized at $\left(\sigma_i', \hat{s}_{-i}\right)$ implies that $v\left(s_i', \hat{s}_{-i}\right) = v\left(\sigma_i', \hat{s}_{-i}\right)$ for all $s_i' \in C\left(\sigma_i'\right)$. Thus the profile $\left(s_i', \hat{s}_{-i}\right)$ is an ideal norm for any $s_i' \in C\left(\sigma_i'\right)$. Pick an $s_i' \in C\left(\sigma_i'\right)$. If s_i' is undominated, then $\left(s_i', \hat{s}_{-i}\right)$ 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 $\left(s_i'', \hat{s}_{-i}\right)$ is an ideal norm for any $s_i'' \in C\left(\sigma_i''\right)$. Thus, either $\left(s_i'', \hat{s}_{-i}\right)$ 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 $\left(s_i^*, \hat{s}_{-i}\right)$ is an ideal norm. Iterating for each player i completes the construction of a pure ideal norm.

Proof of Proposition 2

The utility associated with s_1 is $u_1(s_1) = s_1 - \alpha\delta_1 \left|2s_1 - 10\right|$ 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 $\delta_1 > 1/2\alpha$.

Proof of Proposition 3

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.

Proof of Proposition 5

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.

Proof of Proposition 7

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

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

Solving for δ_1 yields (ii).

Proof of Proposition 8

As before, $\hat{V}(U, P) = 0$, but in the Involuntary Ultimatum situation $\hat{V}(U, A) = 10 - 6\alpha > 0$ 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.

Proof of Proposition 10

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

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

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.

Proof of Proposition 11

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

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

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 $\alpha < 5/7$.

Appendix B: 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.

5.1 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

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

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

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

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

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

Inspecting the latter expression yields the result:

Proposition 12 *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 13 *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 14 *Suppose the exit choice is seen as a choice between social situations, not within. Then,*

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

Proof. Available upon request. ■