TASTES FOR DESERT AND PLACATION: A REFERENCE POINT-DEPENDENT MODEL OF SOCIAL PREFERENCES

Daniel L. Chen*

Abstract I propose a model of behavior in social interactions where individuals maximize a three-term utility function: a conventional consumption utility term and two "social" terms that capture social preference. One social term is a taste for desert, which is maximized when the individual believes the other person is getting what they deserve. The second social term measures the target individuals' anger or gratitude from the interaction which is determined by a value function derived from prospect theory. After introducing the model and generating a series of comparative statics results and derived predictions, I report the results of a series of quasi-field experiments on social preferences. I discuss how the model explains several paradoxes of empirical moral philosophy that are less explicable by current economic models of social preference focusing on outcomes and intentions.

Keywords: Reference points, social preferences, just desert

JEL codes: D6, K2

*Daniel L. Chen, daniel.chen@iast.fr, Toulouse School of Economics, Institute for Advanced Study in Toulouse, University of Toulouse Capitole, Toulouse, France; dchen@law.harvard.edu, LWP, Harvard Law School. First draft: January 2009. Current draft: March 2018. Latest version available at: http://nber.org/~dlchen/papers/Tastes_for_Desert_and_Placation.pdf. Work on this project was conducted while Daniel Chen received financial support from the European Research Council (Grant No. 614708), Swiss National Science Foundation (Grant Nos. 100018-152678 and 106014-150820), Ewing Marion Kauffman Foundation, Institute for Humane Studies, Petrie-Flom Center, Templeton Foundation (Grant No. 22420), Agence Nationale de la Recherche, Berkman Center, Summer Academic Fellowship, and John M. Olin Center for Law, Economics, and Business at Harvard Law School.

1 Introduction

The idea that other's expectations are at the heart of conceptions of fairness can shed light on some of the "paradoxes" of empirical moral philosophy. Consider the classic moral thought experiment where you have the choice whether to pull a lever and divert an out-of-control train careening towards a group of people, but at the cost of killing a strictly smaller number of people standing on the other track: many people find it acceptable to pull the lever — but in a related scenario, few people find it acceptable to push a portly companion into the train's path even if it has strictly better utilitarian outcomes.

One difference between the train diversion and the shove is about expectations. The reason we find it morally acceptable to divert the train is that at the moment when we realize the danger, our expectations for either group now include a high probability of death, with the exact group in danger being a hidden move by nature. We find it objectionable to push the man because the new train danger did not change the expectations for the portly man whose expectation of death did not rise unless there was some unforeseen, purposeful action by another person to put him at risk. The human agency involved in increasing the portly man's risks when he was "spared" by nature seems deeply unfair.

If we stay with the train example for a moment, suppose the larger group was warned to stay off the tracks, were drunk, were trespassing etc., while the victims of the deserted trains were authorized to be on the tracks and were conducting repairs? As we contextualize the problem, our mental calculations about what other people should reasonably expect to happen to them changes and problems that were formally cast as utilitarian, outcome based moral decisions become process-oriented discussions of desert.

Extensive experimental research shows that individuals making decisions, at least in a laboratory context, are not strictly concerned with their own consumption. Departures from purely self-interested behavior, such as donations in public goods games and offers and rejections in ultimatum games are common-place in the laboratory. While it is clear that "social preferences" exist, it is not clear where these preferences come from, how stable they are, how they are distributed among the population or even how they should be modeled.

A large collection of experimental findings in economics has been attributed to perceptions of fairness. However, as Camerer (2003) points out, a weakness of fairness-based explanations is that they typically don't address the question of where fairness preferences might come from. One possibility is social conditioning. Roth et al. (1991) attribute to cultural differences the small, significant differences found between Tokyo, Pittsburgh, and Jerusalem in economic games. Roth (1995) specifies that differences in what is perceived as "fair" or "expected" could explain differences found between American and Israeli proposers in bargaining games. More broadly, Henrich (2000) and Henrich et al. (2001) suggest that economic decisions and reasoning may be heavily influenced by cultural differences, defined as the socially transmitted rules about how to behave. It has been argued that economic theory needs to integrate cultural or moral forces to explain empirical findings.

Even the simplest experimental games designed to analyze social preferences in the laboratory generate complex behavior that cannot be parsimoniously explained by a single, simple model (Charness and Rabin 2002). In experiments, subjects have shown a willingness to punish free-riders, reduce inequality, increase efficiency and enforce social norms, often at substantial cost to themselves. Despite these regularities, there is a non-trivial share of subjects who play laboratory games according to the selfish, homo economicus predictions.

To account for these heterogeneous outcomes, previous models of fairness have proposed that *some* humans are "hard-wired" for reciprocity, inequity aversion, efficiency or other kinds of behavior while others lack other-regarding preferences. Although these modeling approaches lead to simple, tractable models, they seem insufficiently rich to capture the complexity of human behavior. Rather than assume a particular kind of preference, this paper proposes that most social preferences can be understood as manifestations of a more basic human taste for desert — the moral-philosophical notion that people should get what they deserve and further, that for self-interested reasons, humans generally try to satisfy others' expectations about what they deserve, which can be called a preference for placation.

A preference for placation is related to guilt aversion (Battigalli and Dufwenberg 2007): the prototypical cause of guilt would be the infliction of harm or distress on the recipient. For example, in the dictator game, if the game partner expects to receive a certain

amount, behaving selfishly causes feelings of guilt due to not fulfilling the partner's expectations (Battigalli et al. 2013). This definition of guilt is motivated by psychological research looking at a social relationship to a partner (Baumeister et al. 1994). See also Geanakoplos et al. (1989) on incorporating the fulfillment of other's beliefs into a social game. In contrast, a preference for desert is related to notions of distributive justice, which could but does not necessarily include others' expectations (Elizabeth Hoffman 1985; Konow 2000; Gill and Stone 2010; Adams 1966; Rabin 1998).

Desert and placation preferences can explain heterogeneous behavior across different contexts without appealing to the notion of different "types" of people; different outcomes can be traced to different beliefs about how people should play a certain game and accordingly, what they expect from the game.[1] This flexibility might seem to create a vacuous, tautological theory since it is difficult to imagine how one can determine expectations, or even more troubling, beliefs about other's expectations. While it is true that determining expectations is problematic, we can exogenously manipulate expectations by providing new information and then observing whether or not individuals respond in the predicted direction. Further, given the role that social norms play in regulating human behavior through social interactions and the obvious cross-cultural pliability of those norms, it seems that a level of generality and abstraction is needed for the model to organize disparate observations. Much like Koszegi and Rabin (2006), I posit a reference-dependent kind of utility but remain agnostic about how these reference points are formed.

I propose a reference-point dependent model of social behavior where individuals maximize a three-term utility function: a consumption utility term and two "social" terms. One social term captures a preference for desert (others getting what we think they deserve) and the other term a preference for the satisfaction of other's expectations, or to placate them (i.e. them getting what we think they think they deserve). After motivating the modeling assumptions with findings from empirical moral philosophy and evolutionary psychology, I introduce the model and generate some simple comparative statics results, which I then test

[1]To be sure, one might be open to having both different types and different expectations and leave it as an empirical question whether heterogeneity in behavior comes from different types or different expectations.

with experiments using MTurk.

My model is closest to a recent paper by Smith and Wilson (2015), which argues that the need for acceptance in one's group drives behavior previously interpreted as inequity aversion. Smith and Wilson (2015) quote Adam Smith, who believed that an individual makes moral judgments on the propriety of own and other action given the context, but over time the normative rules of propriety change by group consent.

> "Were it possible that a human creature could grow up to manhood in some solitary place, without any communication with his own species, he could no more think of his own character, of the propriety or demerit of his own sentiments and conduct, of the beauty or deformity of his own mind, than of the beauty or deformity of his own face... Bring him into society, and he is immediately provided with the mirror he wanted before. It is placed in the countenance and behavior of those he lives with, which always mark when they enter into, and when they disapprove of his sentiments; and it is here that he first views the propriety and impropriety of his passions, the beauty and deformity of his own mind."

Smith and Wilson (2015) further quote Adam Smith, who said that "Man has a 'love of praise and of praise-worthiness' and a 'dread of blame and blameworthiness', and '[t]he love of praise-worthiness is by no means derived altogether from the love of praise....though they resemble one another...[and]...are connected..., [they] are yet, in many respects, distinct and independent of one another' (Smith 1761)." Moral blame and blameworthiness is like a preference for desert (others getting what we think they deserve). Moral praise and praise-worthiness is like a preference for the satisfaction of other's expectations, or to placate them (i.e. them getting what we think they think they deserve).

Smith and Wilson (2015) proposes an additive interaction model of choice determined by whether an action deserves social praise, whether it is praise-worthy, or both. A multi-stage experiment is used to illustrate principles of fairness. Smith and Wilson (2015) distinguishes from the prior literature on other-regarding behavior defined over own and other's reward payoffs by stating "*when a key prediction of a theory fails, all of its assumptions must be on the table for reconsideration, and the search for a resolution must not exclude consideration of*

entirely different ways of thinking, representing, and modeling the phenomena." My approach takes a middle ground. It builds on the assumptions of other-regarding behavior defined over own and other's reward payoffs, yet it distinguishes from Smith and Wilson (2015) by paying formal attention to reference points in the model and an experiment that exogenously and saliently shifts reference points.

2 Background

Recent work in experimental moral psychology by Greene et al. (2004) highlights the two distinct ways in which humans react to moral dilemmas. The work focuses on the distinction people make between "personal" moral dilemmas that involve an individual causing direct bodily harm to another through their own agency and "impersonal" moral dilemmas that require abstract, utilitarian style reasoning. My model does not involve bodily harm, but it is motivated by the same insight, which is that there is duality in our moral reasoning that can be traced back to evolutionary features of the mind:

> "Evidence from observations of great apes suggests that our common ancestors lived intensely social lives guided by emotions such as empathy, anger, gratitude, jealousy, joy, love and a sense of fairness, and all in the apparent absence of moral *reasoning*. Thus, from an evolutionary standpoint, it would seem strange if human behavior were not driven in part by domain-specific social-emotional dispositions. At the same time, however, humans appear to possess a domain-general capacity for sophisticated abstract reasoning, and it would be surprising as well if this capacity played no role in human moral judgement." (Greene et al., 2004)

Ethologists studying social animals have found that dominance hierarchies are widespread, with those at the top of the hierarchy receiving more food and more resources. One possible evolutionary advantage of the dominance hierarchy is that it makes all animals' reasonable expectations publicly known and prevents socially destructive fighting over each new resource distribution problem that arises. The basic idea is quite simple: we may treat people fairly because we do not want to anger them, and since individuals are "prospectors," we must learn about their expectations and not just their levels of consumption in order to avoid their wrath. In the animal world, there is considerable evidence that fighting among social

animals occurs where expectations (such as who should get some piece of food) are unclear:

> "Baboons ordinarily forage like flocks of birds, fanning out in a search for small vegetable items that are picked off the ground and eaten quickly. The troop members seldom challenge each one another under these circumstances. But when a clump of grass shoots is discovered in elephant dung, or a small animal is killed, the baboons threaten one another and may even fight over the food."[2]

This example also suggests that in circumstances where sharing or allowing others to get what they deserve is costly and in terms of foregone consumption (e.g. small dead animals or grass shoots), self-interest "wins" over the social preferences.

As a rough approximation, it seems the desert utility — the pleasure from seeing others getting what they deserve, which might include efficient outcomes or even justified punishment — is a higher order, more abstract feature of the mind, while placating utility is a more basic, "low-level" feature of the mind since it has such obvious importance for any social animal including our pre-reasoning ancestors.[3] Given insights from prospect theory (Kahneman and Tversky 1979) and the insights from the sociology literature on comparison theory, formulating other-regarding preferences in terms of reference points seems well motivated by the facts. Shaw et al. (2011) finding that combining incentives with social comparisons was most effective in incentivizing individuals in a field experiment is consistent with the relevance of reference points.

Existing models of fairness can be roughly classified by how they treat intentions.

[2] *Sociobiology*, Wilson (2000), p249.

[3] We might also divide this between Type I thinking (also referred to as automatic, cognitive, unconscious) or Type II thinking (reflective, motivational, conscious) (Kahneman 2011). A large collection of findings on the malleability of moral reasoning by judges has been documented in U.S. federal circuit judges (Chen 2017b; Chen et al. 2016), federal district judges (Chen 2017a; Barry et al. 2016), immigration judges (Chen et al. 2016), sentencing judges (Chen and Prescott 2016), and juvenile judges (Eren and Mocan 2016). Some of these findings can be attributed to snap judgments whether from analysis of the first three seconds of oral arguments (Chen et al. 2016; Chen et al. 2017) or from early predictability of judicial decisions based on race or nationality (Chen et al. 2016; Chen and Eagel 2016). One way to model differences in judges' decision-making is through shifts in their reference points about what is the just and fair decision given the circumstances. These reference points may shift consciously or unconsciously. In most of these examples, the desert term matters mostly for a just and fair decision, but in some cases, like asylum decisions, the defendant reference points may play a role. More broadly, when individuals feel they are not being treated justly or fairly, the perceived legitimacy of legal institutions is affected (Chen 2017c).

Models of fairness that do not include intentions, such as Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) cannot explain experimental evidence that shows the importance of motives behind subjective assessments of fairness (Falk et al. 2008). Models that do include intentions (Rabin 1993; Falk and Fischbacher 2006) cannot explain the substantial evidence that people will bear costs to punish others when they have no direct stake or were not harmed directly. For example, third-parties will enforce social norms about sharing or fairness even if it is costly for them to do so (Fehr and Fischbacher 2004) and evidence from neuro-economics suggests that enforcing social norms gives people some kind of pleasure, or at the very least, not enforcing social norms causes displeasure (Spitzer et al. 2007). In addition to a substantial body of experimental economics documenting social preferences, some experiments have focused directly on desert: Eckel and Grossman (1996) conducted an anonymous dictator-game experiment that showed distributions by the dictator became less selfish as the target was seen as more deserving.

The great advantage of a reference-dependent fairness model is that it can accommodate the diversity of human behavior in different contexts as different beliefs about reference points without partitioning people as "types" as is done in intention-free models of fairness (Fehr and Schmidt 1999) or assuming aversion to particular distributions of outcomes. The finding that, for example, business majors and economics students play certain economic games differently seem unlikely to be due to the different mental machinery of those individuals, but rather the context, including their knowledge of the "rational" way to play the game changes their beliefs about what is deserved.[4] A process of updating expectations and reference points during play can parsimoniously explain the findings of Camerer and Fehr (2006) that certain "types" can tip games toward the rational, Nash Equilibrium predictions or the cooperative predictions and provide a mechanism.

In my model of fairness with three components, the consumption-based utility term is a standard economic utility function with the usual properties, the desert-utility function is concave and only reaches a maximum when the other person gets what they "deserve," and

[4]Similarly, the findings that judges decide cases differently and sentence more harshly after economics training can be due to shifts in expectations (Ash et al. 2016).

the backlash-aversion / social capital investment function is the target person's prospect-theory derived value function, centered at the target's subjective reference point. The cost of introducing a reference-dependent theory is complexity and the inability to make predictions without some level of knowledge about the particular context in which behavior is occurring. Koszegi and Rabin (2006), in their model of individual reference-dependent utilities, notes that knowing expectations is difficult in the case of one person — adding another dimension in which one person's beliefs about other's expectations (and their beliefs about the other's beliefs) certainly does not make things simpler. However, simple and wrong is no virtue and it seems that existing simple models cannot account for all the readily apparent stylized facts generated from even the most basic social interactions.

In Eckel and Grossman (1996), distributions by the dictator were affected by whether the target was seen as more deserving. What is novel is fitting this taste for desert into a framework sufficiently flexible to allow selfish behavior, altruism, reciprocity and even spiteful behavior. This can be accomplished by assuming the better angel of wanting to see justice served does not always win, but rather is one consideration, with the other considerations being one's own consumption and the possibility that the other person might retaliate or help you in the future, depending upon whether your actions fell below or above their own subjective expectation. In this paper, I model concern for how the other person will react to some action in terms of a Prospect Theory-derived value function (Kahneman and Tversky 1979). If humans perceive angering others (whose anger or happiness is reference dependent) has a real cost, then we would expect social preferences to consider the expectations of other social actors.

It is obviously not possible to directly observe the evolutionary environment that shaped human reasoning and therefore it is speculative to appeal to evolutionary forces to explain some phenomena. The lack of observability of the ancestral environment or even a fossil record of the evolution of "mental organs" threatens to turn any evolutionary explanation for some feature of the human mind into a "just so" story. However, it is useful to think about the contours of the evolutionary environment and consider some possible reasons why our social mental machinery takes the form that it does.

If loss-aversion or more generally, reference-dependent utilities are a fundamental feature of the primate mind (Chen et al. 2006), it seems probable that the evolution of the human mind — particularly the mental machinery for acting in social situations — was shaped by this fact. Since reference points determine whether our social actions are perceived as gifts, or affronts and these judgements could have consequences relevant to our own well-being, it stands to reason that assessing other's true reference points was adaptive.

Whether an action is interpreted as an affront or a gift will depend, in part, on the skill with which we can estimate the other person's true, subjective expectation from some social situation and adjust our behavior accordingly. For example, judges are more lenient when sentencing defendants on their birthday (Chen and Philippe 2017), and this would be interpreted as a gift by the defendant who does not expect the judge to be lenient due to societal expectations of being gentle on someone's birthday.[5] We are probably also cognizant of the fact that individuals might benefit from over-stating their expectations (and, since self-delusion is probably the best way to present a realistic portrayal of over-inflated expectations, believing their inflated sense of what they deserve, we almost certainly do not take others' self-statements of expectations at face value, but rather make our own assessments of what is fair. Both to prevent ourselves from being cheated or inadvertently angering others or inadvertently giving too generous a gift, it was probably adaptive for our ancestors to develop the ability to quickly assess a social situation and determine what each person deserves and what each person was expecting.

Why should we care what others subjectively feel about a transfer, especially if we have some true altruistic motive (in which case their actual material well-being is important while their subjective, frame-dependent perception is irrelevant)? The answer is that an aggrieved party might retaliate with violence and future non-cooperation. If fear of retaliation

[5]I use expectation or reference point in a very general sense and not in the strict mathematical sense; an expectation might mean that the other person follows a certain custom or norm. Outside the lab, conceptions of human rights may also hinge on the context, for example, on rights pertaining to asymmetric virginity premiums (Chen 2005) or sexual harassment (Chen and Sethi 2016) or for repugnance norms related to right of free speech (Chen 2015a) and abortion (Chen et al. 2017). The malleability of injunctive norms to formal institutions such as the law (Chen and Yeh 2016, 2014) or markets (Chen 2015b; Chen and Lind 2016; Chen 2016) is suggestive of the relevance of reference points, but it's hard to know, since many other things change at the same time as formal institutions. This paper shares the experimental approach to measure normative commitments (Chen et al. 2016; Shaw et al. 2011).

can explain other-regarding preferences, why do we need a separate taste for desert? Perhaps the strongest reason is that social behavior is clearly not governed solely by a concern for other's ability to help or hurt us. As noted before, efforts are made to punish free-riders, enforce social norms and promote other social goods like efficiency. Aside from the evidence that requires explanation, another motivation for desert preferences is that humans do seem to incorporate "higher-order," social concerns into their thinking and genuinely suffer a real psychic cost when they see what they perceive as valuable social norms violated, regardless of whether the situation directly impinges upon their narrowly-defined welfare. The relevance of non-consequentialist motivations, such as duty has been suggested in some experiments.[6]

Another answer is that a model needs a taste for desert because it is an empirical regularity. Third-parties will spend resources to punish others, even if the person has done nothing to them (Fehr and Fischbacher 2004). If our only concern was our own consumption and possibility of reprisal, it would never be optimal to punish someone and further, we would never show compassion for those too weak to hurt us or help us — and yet we do, albeit perhaps not as much, ceteris paribus, as the regard we show to those who are more useful to us. Even in simple ultimatum games, without a taste for desert, we cannot explain the rejection of low-offers without assuming some kind of expectation for repeated interactions or, as some have done, a taste for equity. A second reason is that so long as the consumption rewards of defection are not so great, a taste for desert can immediately generate the tit-for-tat strategy without having to assume a special kind of taste for reciprocity (Axelrod and Hamilton 1981). In a repeated game, a person with a taste for desert will respond to a defection with defection in order to bring the defector back to their proper reference point, but once this is done, is once again interested in cooperation.

3 Model

The dictator game has the same relationship to the investigation of social preferences that the fruit fly does to the study of genetics. It is perhaps the simplest possible interaction

[6] For an economic model and test of the categorical imperative, see Chen and Schonger (2016; 2017).

that has a by-design social preference / fairness component. A decision-maker might be a judge or prosecutor who has to determine what is the fair sentencing decision or sentence to charge. Being too harsh or too lenient, from their perspective, is undesirable, while, a defendant experience gains with sentencing leniency.

Since this paper's intention is not to propose a general model for all social interactions, I present a utility function specific to the dictator game. This formulation can serve as the building block for more complicated analysis, such as public goods games, ultimatum games, etc. This game is simple, deterministic and it is easy to specify expectation beliefs (unlike in more complicated settings where expectations might include certain actions or following more complex social norms than simply sharing some quantity of resources). Suppose there are two individuals playing a dictator game, Gabriel the giver and Randy the recipient. Gabriel has to decide the optimal transfer $x \in [0, w]$ to give to Randy. Gabriel's utility from making a transfer is given by equation 1.

Proposer's utility from making a transfer is given by Equation 1.

$$U_g(x) = u(w - x) + u_p(x - x_p^{RP}) + v_r(x - x_r^{RP}) \tag{1}$$

Where w is her initial wealth, x_p^{RP} is her subjective reference point of what is the just transfer in this situation and x_r^{RP} is her best estimate or belief about what the Receiver's reference point is for the game. Gabriel's belief about what Randy deserves (i.e. her belief about his reference point), irrespective of the costs that would be incurred in actually making this transfer, is captured by x_p^{RP}. Randy's actual beliefs about what he deserves are x_r^{RP}.

The model assumptions are:

- The consumption utility term is increasing and is strictly concave: $u'() > 0$, $u''() \leq 0$
- The desert utility function is concave and is maximized at the desert point: $x_p^{RP} = argmax_x u_p(x, x_p^{RP})$
- The $v_r()$ function is concave in gains and is zero when the Receiver receives exactly what Proposer believes he believes he should receive: $v_r(x_r^{RP}) = 0$
- If $x > x_r^{RP}$, then $v_r(x - x_r^{RP}) > 0$ and $v_r''() \leq 0$, otherwise $v_r(x - x_r^{RP}) \leq 0$, $v_r''() \geq 0$.

These assumptions generate some very simple comparative statics results, which are presented below and then tested. For ease of notation, let $x_r^{RP} = x_r$ and $x_p^{RP} = x_p$. For intuition, let each second-derivative be negative so that $u''(w - x^*) + u_p''(x^* - x_p) + v_r''(x^* - x_r) < 0$. After the basic results are presented with a figure, this assumption is relaxed and discussed.

Lemma 1: The optimal transfer x^* is strictly increasing in what the Proposer believes the Receiver deserves, or: $\delta x^*/\delta x_p > 0$.

Proof. The first-order condition for the Proposer's maximization problem is

$$-u'(w - x) + u_p'(x - x_p) + v_r'(x - x_r) = 0$$

This first-order condition defines an implicit relationship between the optimal transfer x^* and the two exogenous reference points: $x^* = x^*(x_p, x_r)$. Differentiating with respect to the Proposer's reference point x_p, with $x_p^*(x_p, x_r)$ representing the partial derivative of the optimal transfer with respect to x_p.

$$u''(w - x^*)x_p^*(x_p, x_r) + u_p''(x^* - x_p)(x_p^*(x_p, x_r) - 1) + v_r''(x^* - x_r)x_p^*(x_p, x_r) = 0$$

Solving, yields:

$$x_p^*(x_p, x_r) = \frac{u_p''(x^* - x_p)}{u''(w - x^*) + u_p''(x^* - x_p) + v_r''(x^* - x_r)}$$

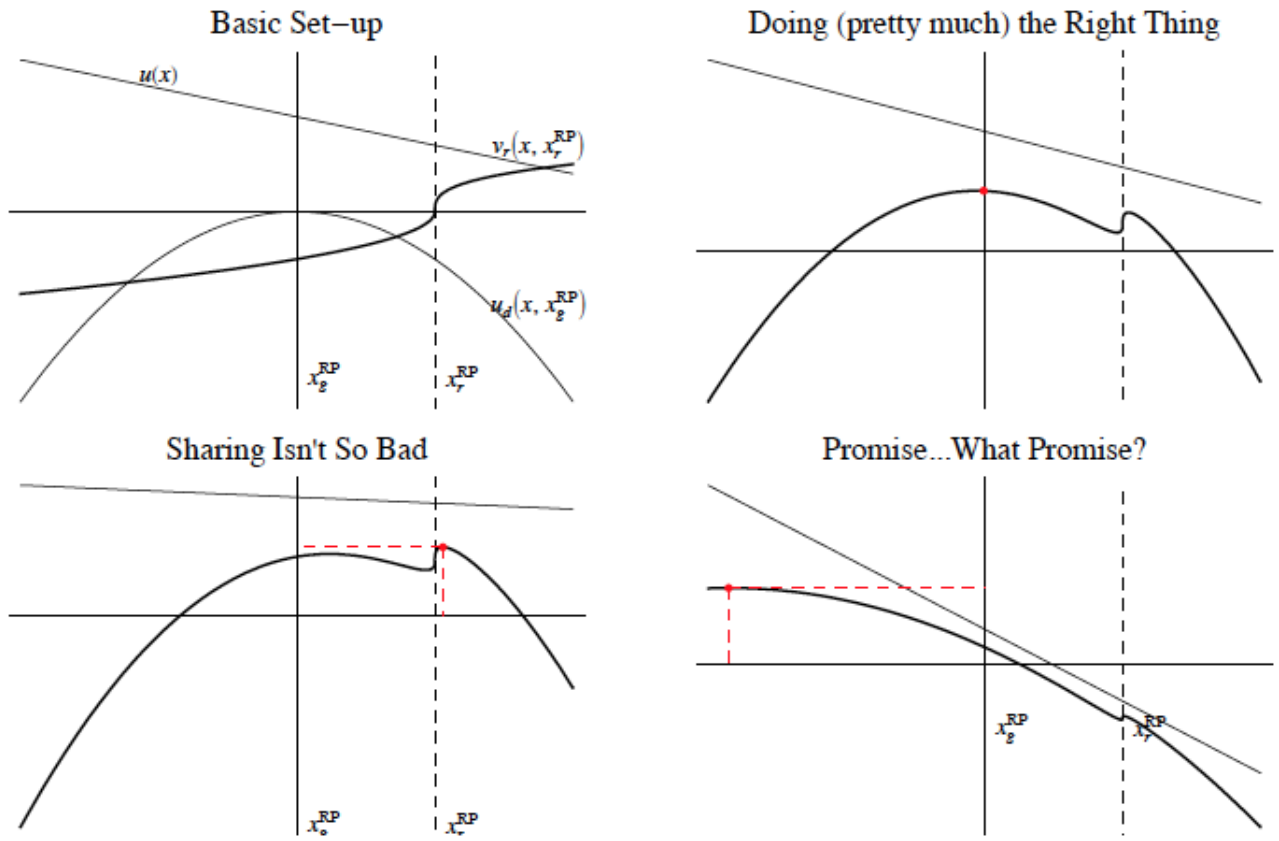
If the second derivative of all of the three terms have a negative sum ($u''(w - x^*) + u_p''(x^* - x_p) + v_r''(x^* - x_r) < 0$), then $x_p^*(x_p, x_r) \geq 0$, regardless of the particular value of x^*.

Lemma 2: The optimal transfer x^* is strictly increasing in what the Proposer believes the Receiver believes he deserves, or: $\delta x^*/\delta x_r > 0$.

Proof. Following the same procedure in the proof of the first lemma,

$$x_r^*(x_p, x_r) = \frac{v_r''(x^* - x_p)}{u''(w - x^*) + u_p''(x^* - x_p) + v_r''(x^* - x_r)}$$

and thus $x_r^*(x_p, x_r) \geq 0$, regardless of the particular value of x^*.



Note: The upper-left panel shows all the component utility terms as well as Gabriel's (the giver) desert reference point for Randy (the receiver) and Randy's own reference point. In the remaining three panels, Gabriel's composite utility $U()$ is shown as well as her utility-maximzing transfer (a red dot). In the upper-right panel, the desert term dominates and Gabriel transfers what she believes Randy deserves according to her own estimate. From Randy's perspective, this action looks like a punishment and he is angry. In the lower-left panel, the slop of the consumption utility is smaller and hence Gabriel transfers a "gift" (from her perspective) to Randy, which he finds to be slightly in the domain of gains. In the lower-right panel, the slope of the consumption utility is just too steep and Gabriel cannot resist cheating Randy and giving him less than even she thinks he deserves.

These lemmas are illustrated in the accompanying figure.

The figure shows that the Receiver is loss averse relative to his reference point. If $x < x_r^{RP}$, then $v_r(x - x_r^{RP}) \leq 0$, $v_r''() \geq 0$, and the proof is slightly more complicated.

Loss-Averse Receiver:

If the second derivative of $v_r()$ is positive enough, then $u''(w-x^*)+u_p''(x^*-x_p)+v_r''(x^*-x_r) > 0$ can lead to the opposite conclusion for Lemma 1. The intuition can be illustrated by supposing the Proposer is in the lower-left quadrant of the figure ("Sharing Isn't So Bad"), and has made a decision at the Recipient's reference point. Now, the optimal decision can decrease as the Recipient's reference point increases. This is because the Proposer has a concave cost of deviating from the optimal decision. As the Recipient's reference point increases and $v_r()$ shifts rightward, the Proposer may decide "what-the-hell" and jump to a lower optimum (in

the upper-right quadrant of the figure) rather than placate the Recipient. This is consistent with situations when the Recipient asks or expects too much, leading the Proposer to ignore his expectations. In the figure, the marginal cost of meeting the Recipient's expectations increases with pressure from the desert and consumption utility terms.

Popular terms like the "What the Hell Effect" (Ariely 2012; Baumeister and Heatherton 1996) capture this behavioral tendency—the cognitive cost of deviating downwards in a large scale is not much larger than in a small scale. A number of theoretical papers show that the curvature of bliss-point deviations has important implications for decision making in social and political settings. For instance, as discussed by Osborne (1995) and shown by Kamada and Kojima (2014), concavity drives polarization in political platforms. This is since voters do not perceive a difference between a policy slightly away from their bliss point and a policy far away. Hence, unless a candidate adheres very closely to a group of voters' preferences, these voters will not vote at all, implying that in a polarized electorate political platforms will be polarized when ideological costs are concave but not otherwise. In a different setting, Chen et al. (2016) shows that concavity in deviating from a bliss point can lead judges to cave-in – dissent less often – than other judges despite having the least say in shaping court decisions.

In Lemma 2, if the Proposer's reference point decreases, the optimal decision may jump to a lower optimum. This is because the numerator, $v_r''() \geq 0$, so if the Receiver is loss averse enough and $u''(w-x^*)+u_p''(x^*-x_p)+v_r''(x^*-x_r) > 0$, the original prediction holds. The intuition can be seen in the lower-left quadrant of the figure. As the Proposer's reference point decreases and $u_p(x, x_p^{RP})$ shifts leftward, the Proposer's utility at x_r^{RP} decreases more sharply than the utility for some decision less than x_r^{RP}. That is, the marginal cost of meeting the Recipient's expectations increases with pressure from the desert term. Thus, if the Recipient is sufficiently loss averse, decreasing the reference point of the Proposer leads the Proposer to decide "what-the-hell" and jump to a lower optimum.

Example:

An application of the model is that we might score judges on their indifference to others. One way to know if a judge is indifferent, is by observing where behavioral biases

arise. Psychologists find many effects of moderate sizes in the lab, so settings where people are closer to indifference among options are more likely to lead to detectable effects outside of it. The flatter is the judge's parabola (in the figure), the more sensitive the judge is to behavioral biases. Another way to put it is that the judge becomes less predictable.

4 Experimental Design & Results

To test the comparative static results, I use a contextualized dictator game and a contextualized gift game in an MTurk experiment. Subjects were recruited from an online temporary work environment that allows users of the website to complete simple tasks proposed by others in exchange for payment. In this real market, workers perform tasks that are generally hard for computers but easy for humans. Common tasks include image tagging (i.e. writing captions for images), categorizing websites by their context, transcribing audio files to text, transcribing badly scanned text documents or extracting particular pieces of information from scanned documents. In the experiments, I presented a task in which subjects were asked to rate the work of another worker. Using other workers to check the work of others is a very common task on the website.

The experimental task was a real-work task (though not very taxing): subjects were shown a scanned image of text and the transcription work of "another worker." Subjects were asked to compare the scanned image to the transcription and then complete some additional task, depending on the particular experiment. In the pure gift games, subjects were asked to assess a penalty or bonus for the work, while in the dictator games, subjects were asked to split a bonus with the original "worker." On the website, awarding bonuses for good work is commonplace, as is using multiple workers to check the work of each other, so it is unlikely that any worker found the experiments unusual or artificial.

Although de-contextualized experiments have been the norm is experimental economics, it is possible that typical laboratory scenarios never really strip away context and create a "pure" testing environment; many experimental protocols can evoke a very specific context with well-developed norms, namely that of a true, competitive game.[7] While having

[7] If your mother-in-law lands on "Boardwalk" when playing Monopoly, you will charge her the full price; when she comes to visit and stays at your guest-bedroom, she probably will not be billed.

a context certainly does not make these experiments better than laboratory experiments, it also does not seem like a handicap, especially since the experiment was in a real work environment with subjects unaware that they were participating in an experiment. Further, subjects were randomly assigned to treatment and control upon accepting the work and subjects can not interact with each other in anyway.

To recruit subjects, I placed tasks on the site and advertised \$0.12 - \$0.15 for completing the task.[8] Once workers accepted, they would click on a link that would re-direct them to a script that randomly assigned subjects to treatment groups (basically by using a series of if-then statements with URL re-directs). The study was conducted on April 2, 2009.

To give a sense of the stakes, in experiments involving data entry of the text image (Chen 2016; Chen and Horton 2016), a paragraph takes about 100 seconds to enter so a payment of \$0.10 per paragraph is equivalent to \$86.40 per day. The current federal minimum wage in the Unites States is \$58/day. In India, payment rate depends on the type of work done, although the "floor" for data entry positions appears to be about \$6.38/day.[9] Most of the subjects came from the U.S. or India (I conducted pooled analysis throughout). In other studies conducted by the author using this data entry task, subjects did seriously enter the text image. In one study, one worker emailed saying that \$0.10 was too high and that the typical payment for this sort of data entry was \$0.03 cents per paragraph. The decision-task involves reading (not transcribing) a single paragraph and making 1 decision, with an additional bonus possible. For instance, \$0.30 is the maximum a subject could obtain and is up to 10 times the expected wage.

5 Placation: Manipulate Receiver Reference Point x_r^{RP}

Lemma 2 predicts that the proposer's transfer will increase if the proposer believes that the receiver has a higher reference point. In this experiment, subjects were randomized into two groups. In one group, subjects were told that performing a transcription received a

[8] Prices varied across experiments but not within experiments since some tasks are more onerous than others and therefore would require a higher payment to attract subjects. While there is non-random selection into the experiment (which is the case with all experiments), there was still random assignment into the treatment groups.

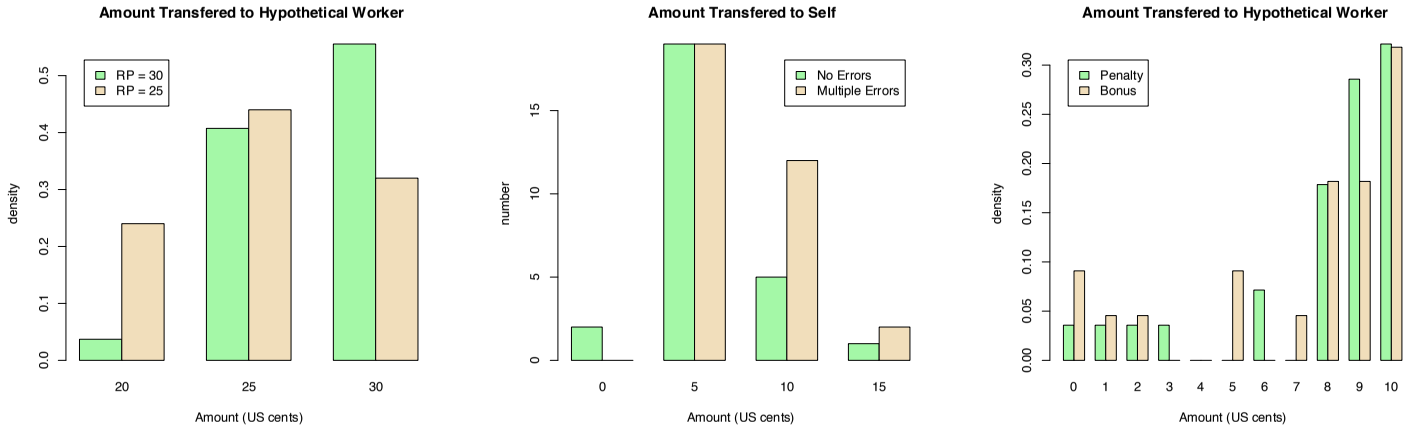
[9] Payscale, Salary Snapshot for Data Entry Operator Jobs, http://www.payscale.com/research/IN/Job=Data_Entry-_Operator/Salary?, accessed June 17, 2011.

base payment of \$0.25 but that they were eligible for a \$0.05 bonus, a \$0.05 penalty or they could be kept at the status quo of \$0.25. Here, 11 subjects awarded \$0.25, 6 awarded \$0.20, and 8 awarded \$0.30. In the second group, subjects were told that the base rate was \$0.30 and that they could assess a \$0.05 penalty or a \$0.10 penalty or keep the status quo. In this treatment, 15 subjects chose to award \$0.30, 11 awarded \$0.25, and 1 awarded \$0.20. Note that in both groups, the possible outcomes for the worker are the same: $x = \{20, 25, 30\}$. Figure 1a plots the effect that the reference point had on transfers. Table 1 reports the results of a means comparison test of transfers in the two groups. It can reject the null-hypothesis of invariance to reference-points ($p < 0.05$).

6 Desert: Manipulate Proposer Reference Point x_p^{RP}

In both groups, subjects inspect the transcription work, which they can compare to the scanned image of the work pasted directly above the text box containing the work. The transcription sentences are numbered and subjects are asked to check a box indicating whether or not the line contained an error. After assessing the work, subjects are asked to split a \$0.15 bonus between themselves and the other worker in \$0.05 increments. To ensure there was more variation in outcomes, the 50-50 split option was not present. This also encourages subjects to think harder about their splitting decision in light of their task with error-checking. Figure 1b shows the distribution of transfers for each group. In the treatment group with no errors, 2 subjects transferred \$0.00 to themselves, 19 transferred \$0.05 to themselves, 5 transferred \$0.10 to themselves, and 1 transferred all \$0.15 to themselves. In the treatment group with multiple errors, 0 subjects transferred \$0.00 to themselves, 19 transferred \$0.05 to themselves, 11 transferred \$0.10 to themselves, and 2 transferred \$0.15 to themselves.

Comparing group means, we see in Table 2 that the group with the higher number of errors does have a higher average self-transfer ($p < 0.1$). Table 3 shows that the number of errors found by the subjects is highly correlated with the number of true latent errors in the treatment and control transcriptions ($p < 0.001$). Table 4 shows that the amount of money subjects transferred to themselves is strongly correlated with number of errors found



(a) Transfers Increasing in x_r^{RP} (b) Transfers Increasing in x_p^{RP} (c) Lack of Penalty vs. Bonus Frame

Figure 1: Panel 1(a) shows the results of an experiment in which subjects inspected a transcription by a hypothetical worker and then assigned a bonus or penalty. In the RP=25 group, subjects were told the worker was paid a base rate of 25, which they could modify with a bonus of 5 cents or a penalty of 5 cents. In the RP=30 group, subjects were told the worker was paid a base rate of 30, which they could modify with a penalty of 5 cents or a penalty of 10 cents. Panel 1(b) shows the results on an experiment testing whether subjects would award themselves more money and a hypothetical worker less money if they perceived the other worker as having made more errors. Subjects were randomly assigned to the "No Errors" group or the "Multiple Errors" (3 obvious spelling mistakes) group. In both groups, subjects were asked to examine a transcription prepared by a fellow worker and marked check-boxes indicating whether a given line contained an error. After this task, subjects were asked to split a 15 cent bonus between themselves and the other worker, with the allowable distributions being a self-award of 0, 5, 10 or 15 cents. Panel 1(c) shows the results of an auxiliary experiment designed to test whether subjects perceive of penalties and bonuses independent of the effects those actions have. In the experiment, subjects were randomized to a treatment (N=22) and control (N=28) groups: in the treatment group, subjects were asked to award a bonus from $[0, 10]$, while in the control, subjects were asked to assess a penalty from $[0, 10]$.

Table 1: Means Comparisons

	p-value	30 Cent RP	25 Cent RP
H0: No difference in means	0.02	27.59	25.40

Table 2: Means Comparison Test: Self-Transfers

	p-value	Multiple Errors	No Errors
H0: No difference in means	0.07	7.42	5.93

($p < 0.001$) but not with any other subject characteristics.

One criticism is that the number of errors might be endogenous, with cognitively-impaired people missing errors or greedy people inflating the number of errors to justify taking more. To control for this possibility, Table 5 includes a regressor for the number of wrong answers (the difference between the true number of errors and the number they reported). The table shows that the number of incorrectly identified mistakes (the coefficient on "wrong") does not have an effect on transfers: in other words, it does not appear that subjects were artificially inflating the number of mistakes to justify a greater transfer.

One feature that is interesting is that assessing the high-error work requires only a marginally greater amount of work but leads to a significant decrease in bonuses. It does not seem likely that subjects feel that they deserve more for having to examine the error-filled work (they are not asked to correct the work or even identify precisely what the errors are). It is far more likely that they perceive the worker making many errors undeserving of a full bonus and are willing to benefit from the situation, but they are unwilling to extract the full bonus when they perceive the worker as having done perfect work (in fact, not one worker took the full bonus in the perfect transcription group).

7 Penalty vs. Bonus Framing Effects

One possible concern with experiments that require subjects to assign penalties vs.

Table 3: Found Errors vs. Actual Errors

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5556	0.1655	3.36	0.0014
w	1.2929	0.2232	5.79	0.0000

Table 4: Self-Transfer vs. Found Errors and Other Covariates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.2395	1.7256	1.88	0.0659
age	0.0187	0.0380	0.49	0.6247
us	0.7901	1.1218	0.70	0.4843
male	1.1137	0.8876	1.25	0.2150
errors	1.7385	0.4277	4.06	0.0002
w	−0.8132	0.9227	−0.88	0.3820

Table 5: Self-Transfer vs. Actual Errors and Incorrectly Identified Errors

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.2670	1.8086	1.81	0.0765
age	0.0195	0.0410	0.48	0.6358
us	0.7635	1.2265	0.62	0.5363
male	1.1024	0.9178	1.20	0.2350
errors	1.7165	0.5819	2.95	0.0047
w	−0.7593	1.3338	−0.57	0.5716
wrong	−0.0369	0.6537	−0.06	0.9552

bonuses is that, even without a reference-point model, subjects might view penalties and bonuses as fundamentally different. To an extent, the meanings of the words penalties and bonuses creates a frame, though they both imply some action that was unexpected.

In a third experiment, subjects were randomly assigned to two groups: in the first group, subjects were asked to assess transcription work and assign a penalty, while in the second group, subjects were asked to assess a bonus between \$0.00 and \$0.10. Both groups examined the same transcription. The transcription had two errors — "principals" was spelled "principalss' and "decision" is spelled "dicision." Other than these errors, the transcription is accurate. In one group, subjects were asked to assess a penalty between \$0.00 and \$0.10, while in the other, subjects were asked to assess a bonus between \$0.00 and \$0.10. Critically, subjects were not informed about the base rate at which the worker was paid and thus was not supplied any information with which to infer the hypothetical worker's reference point.[10]

The original hypothesis was that subjects would be more averse to imposing penalties than to withholding a bonus for poor performance. The data did not, however, bear out this hypothesis. Figure 1c is a histogram of the implied "transfer" i.e. the bonus b or the portion of

[10]To be sure, the action space and range is slightly different from the second experiment.

the penalty not imposed, $10-p$. In the bonus frame, there were 22 subjects and in the penalty frame, there were 28 subjects. What is surprising about this plot is the consistency between the two treatment groups. Both show a marked bi-modality, with most people awarding the maximum transfer and some awarding only a small transfer with very few splitting the difference with payments of 4, 5, or 6.

Although these results are tangential to the theory, they do speak to the inadequacy of assuming that a) people are unconditionally altruistic when they have no stake (why not give the full bonus?) and that b) aversion to putting others below their expectations is not the only factor considered by subjects (if subjects believed that workers had high expectations, then they wouldn't risk not giving full bonuses).

One interpretation for the consistency between the treatment groups is that since the subject can observe the task and the response, the subject can determine the optimal base rate the employer and the hypothetical worker would have signed, with beliefs about the world "filling in" the appropriate level of quality then imputes the contractual arrangement that would have been made. Suppose for example that the average belief about what the worker deserved is \$0.25 compared to \$0.30 for a perfect transcription. In the penalty group, subjects observe the punishment possibilities $[0, 10]$ and assume a contract was written for 30 cents, with the expectation of penalties and therefore, they feel no qualms about assessing a \$0.05 penalty. In contrast, the bonus group has the same beliefs but assumes that the base contract was \$0.20 with the expectation of a bonus between $[0, 10]$, with \$0.10 for perfect work. In this group, they assess a bonus of 5 cents to bring the worker to their "deserved" wage.

8 Subject Motivation

If subjects have social preferences and presumably have some prior belief about Freddy's marginal utility of money and the experimenter's marginal utility of money, any reasonable belief about the general financial situation of requesters and turkers would call for a strong turker-bias. If it is weakly more pleasant to award bonuses than exact penalties, especially when costs to the subject are equal, and we assume that individuals are not sadists

nor prone to random clicking, why isn't the full transfer made in every case?

Part of the story is certainly that subjects perceive of themselves as agents of the principal (the experimenter). However, if the agent even weakly prefers paying bonuses to enacting penalties, not paying a full transfer in the principal-agent context only makes sense if they also perceive themselves as playing a repeated-game with the principal and that somehow "incorrect" ratings will jeopardize a future relationship. While this repeated-games interpretation is possible, the complete absence of promulgated standards in the task description, the advertisements exhorting speed and easiness (not careful consideration) and the generally small stakes make this interpretation questionable.

It is possible that Mturkers are blindly punching keys and making mouse clicks to avoid expending any effort at all, but it seems more likely that most workers are taking their task at least somewhat seriously.

9 Conclusion

Tastes for desert and placation could explain results such as those found by Bohnet and Zeckhauser (2004), which is that people are averse to getting betrayed and will demand a premium for accepting a risk of betrayal above the risk premium that would require for a risky lottery where nature makes the pay-off determination. It is perfectly reasonable if, looking forward, a rational agent with a taste for desert knows that she will feel, ex post, an obligation to punish the cheater if she is betrayed, and further, this punishment is especially painful due to the loss-aversion of the target of the punishment. This observation is similar to the anecdote related in Lazear et al. (2012) about someone who doesn't like to share but doesn't want to *not* share might cross the street to avoid a beggar, a phenomenon that Andreoni et al. (2012) examined in the field.

While not manipulated in these experiments, there is substantial empirical evidence that social distance can affect behavior towards others (Hoffman et al. 1996, 1999; Bohnet and Frey 1999). The authors cited disagree about the mechanism — i.e. whether individuals are worried about retaliation or whether knowing the identity of one's "victim" increases the salience of what is wrong with stiffing the receiver in a social dictator game – but the model here captures both effects.

One seemingly counter-intuitive implication of the model is that we might actually punish more harshly those people who are socially "close". Judges sentence defendants more harshly when they share the same first initial (Chen and Prescott 2016). Kenyan brokers are more likely to defraud members of their own ethnic group (Yenkey 2015). The ratio of murders by relatives to murders by strangers suggests that social distance is no guarantee of strictly pleasant relationships.

While people in many cultures may consider themselves to have capacity for moral judgment—and expect other people to agree with their moral judgments—they may disagree about what constitutes the morally good thing to do in various circumstances. Such disagreements take place both within and across cultures. Different individuals often have different views on what they think is right or just, but where do these ideas come from? This paper proposes that these ideas come from expectations of what is just and fair.

References

Adams, J. S. (1966). *Inequity In Social Exchange*, Volume 2. Academic Press.

Andreoni, J., J. M. Rao, and H. Trachtman (2012, June). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. Working paper.

Ariely, D. (2012). *The (Honest) Truth About Dishonesty*. New York: Harper Collins Publishers.

Ash, E., D. L. Chen, and S. Naidu (2016). The Effect of Conservative Legal Thought on Economic Jurisprudence. Technical report.

Axelrod, R. and W. D. Hamilton (1981). The evolution of cooperation. *Science 211*(4489), 1390–1396.

Barry, N., L. Buchanan, E. Bakhturina, and D. L. Chen (2016). Events Unrelated to Crime Predict Criminal Sentence Length. Technical report.

Battigalli, P., G. Charness, and M. Dufwenberg (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization 93*, 227–232.

Battigalli, P. and M. Dufwenberg (2007, May). Guilt in games. *The American Economic Review 97*(2), 170–176.

Baumeister, R. F. and T. F. Heatherton (1996). Self-Regulation Failure: An Overview. *Psychological Inquiry 7*(1), 1–15.

Baumeister, R. F., A. M. Stillwell, and T. F. Heatherton (1994). Guilt: An interpersonal approach. *Psychological Bulletin 115*(2), 243.

Bohnet, I. and B. S. Frey (1999). Social distance and other-regarding behavior in dictator games: Comment. *The American Economic Review 89*(1), 335–339.

Bohnet, I. and R. Zeckhauser (2004, December). Trust, risk and betrayal. *Journal of Economic Behavior & Organization 55*(4), 467–484.

Bolton, G. E. and A. Ockenfels (2000). Erc: A theory of equity, reciprocity, and competition. *The American Economic Review 90*(1), 166–193.

Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. The Roundtable Series in Behavioral Economics. Princeton University Press.

Camerer, C. and E. Fehr (2006). When does "economic man" dominate social behavior? *Science 311*(5757), 47–52.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics 117*(3), 817–869.

Chen, D., Y. Halberstam, and A. Yu (2017). Covering: Mutable Characteristics and Perceptions of Voice in the U.S. Supreme Court. *Review of Economic Studies*. invited to resubmit, TSE Working Paper No. 16-680.

Chen, D., Y. Halberstam, and A. C. L. Yu (2016, October). Perceived Masculinity Predicts U.S. Supreme

Court Outcomes. *PLOS ONE 11*(10), 1–20. e0164324.

Chen, D. and A. Philippe (2017). Reference points, mental accounting, and social preferences: Sentencing leniency on birthdays. Technical report, mimeo.

Chen, D. L. (2005, November). Gender Violence and the Price of Virginity: Theory and Evidence of Incomplete Marriage Contracts. Working paper, University of Chicago, Mimeo.

Chen, D. L. (2015a, August). Can Markets Overcome Repugnance? Muslim Trade Reponse to Anti-Muhammad Cartoons. Working paper, ETH Zurich, Mimeo.

Chen, D. L. (2015b, Spring). Can markets stimulate rights? On the alienability of legal claims. *RAND Journal of Economics 46*(1), 23–65.

Chen, D. L. (2016, October). Markets, Morality, and Economic Growth: Competition Affects Moral Judgment. TSE Working Paper No. 16-692.

Chen, D. L. (2017a). Mood and the Malleability of Moral Reasoning. TSE Working Paper No. 16-707.

Chen, D. L. (2017b). Priming Ideology: Why Presidential Elections Affect U.S. Judges. *Journal of Law and Economics*. resubmitted, TSE Working Paper No. 16-681.

Chen, D. L. (2017c). The Deterrent Effect of the Death Penalty? Evidence from British Commutations During World War I. *American Economic Review*. resubmitted, TSE Working Paper No. 16-706.

Chen, D. L., X. Cui, L. Shang, and J. Zheng (2016). What Matters: Agreement Among U.S. Courts of Appeals Judges. *Journal of Machine Learning Research*. forthcoming, TSE Working Paper No. 16-747.

Chen, D. L., M. Dunn, R. G. C. D. Costa, B. Jakubowki, and L. Sagun (2016). Early Predictability of Asylum Court Decisions. Technical report.

Chen, D. L. and J. Eagel (2016). Can Machine Learning Help Predict the Outcome of Asylum Adjudications? Technical report.

Chen, D. L. and J. J. Horton (2016, June). Are Online Labor Markets Spot Markets for Tasks? A Field Experiment on the Behavioral Response to Wage Cuts. *Information Systems Research 27*(2), 403–423. TSE Working Paper No. 16-675.

Chen, D. L., V. Levonyan, and S. Yeh (2017). Do Policies Affect Preferences? Evidence from Random Variation in Abortion Jurisprudence. *Journal of Political Economy*. TSE Working Paper No. 16-723, under review.

Chen, D. L. and J. T. Lind (2016, December). The Political Economy of Beliefs: Why Fiscal and Social Conservatives/Liberals (Sometimes) Come Hand-in-Hand. under review, TSE Working Paper No. 16-722.

Chen, D. L., M. Michaeli, and D. Spiro (2016, August). Ideological Perfectionism. TSE Working Paper No. 16-694.

Chen, D. L., T. J. Moskowitz, and K. Shue (2016). Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires. *The Quarterly Journal of Economics 131*(3),

1181–1242.

Chen, D. L. and J. J. Prescott (2016). Implicit egoism in sentencing decisions: First letter name effects with randomly assigned defendants.

Chen, D. L. and M. Schonger (2016). Social Preferences or Sacred Values? Theory and Evidence of Deontological Motivations. *American Economic Journal: Microeconomics*. invited to resubmit, TSE Working Paper No. 16-714.

Chen, D. L. and M. Schonger (2017, March). A Theory of Experiments: Invariance of Equilibrium to the Strategy Method of Elicitation. TSE Working Paper No. 16-724.

Chen, D. L., M. Schonger, and C. Wickens (2016, March). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88 – 97.

Chen, D. L. and J. K. Sethi (2016, August). Insiders, Outsiders, and Involuntary Unemployment: Sexual Harassment Exacerbates Gender Inequality. invited to resubmit, TSE Working Paper No. 16-687.

Chen, D. L. and S. Yeh (2014, August). The Construction of Morals. *Journal of Economic Behavior and Organization 104*, 84–105.

Chen, D. L. and S. Yeh (2016, September). How Do Rights Revolutions Occur? Free Speech and the First Amendment. TSE Working Paper No. 16-705.

Chen, M. K., V. Lakshminarayanan, and L. R. Santos (2006, June). How basic are behavioral biases? evidence from capuchin monkey trading behavior. *The Journal of Political Economy 114*(3), 517–537.

Eckel, C. C. and P. J. Grossman (1996, October). Altruism in anonymous dictator games. *Games and Economic Behavior 16*(2), 181–191.

Elizabeth Hoffman, M. L. S. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *The Journal of Legal Studies 14*(2), 259–297.

Eren, O. and N. Mocan (2016). Emotional judges and unlucky juveniles. Working paper.

Falk, A., E. Fehr, and U. Fischbacher (2008, January). Testing theories of fairness–intentions matter. *Games and Economic Behavior 62*(1), 287–303.

Falk, A. and U. Fischbacher (2006). A theory of reciprocity. *Games and Economic Behavior 54*(2), 293–315.

Fehr, E. and U. Fischbacher (2004). Third-party punishment and social norms. *Evolution and Human Behavior 25*(2), 63–87.

Fehr, E. and K. M. Schmidt (1999, August). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics 114*(3), 817–868.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). Psychological games and sequential rationality. *Games and Economic Behavior 1*(1), 60–79.

Gill, D. and R. Stone (2010). Fairness and desert in tournaments. *Games and Economic Behavior 69*(2), 346–364.

Greene, J. D., L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron 44*(2), 389–400.

Henrich, J. (2000). Does culture matter in economic behavior? ultimatum game bargaining among the machiguenga of the peruvian amazon. *The American Economic Review 90*(4), 973–979.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath (2001, May). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *The American Economic Review 91*(2), 73–78.

Hoffman, E., K. McCabe, and V. L. Smith (1996). Social distance and other-regarding behavior in dictator games. *The American Economic Review 86*(3), 653–660.

Hoffman, E., K. McCabe, and V. L. Smith (1999). Social distance and other-regarding behavior in dictator games: Reply. *The American Economic Review 89*(1), 340–341.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–292.

Kamada, Y. and F. Kojima (2014). Voter Preferences, Polarization, and Electoral Policies. *American Economic Journal: Microeconomics 6*(4), 203–236.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *The American Economic Review 90*(4), 1072–1091.

Koszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics 121*(4), 1133–1165.

Lazear, E. P., U. Malmendier, and R. A. Weber (2012, January). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics 4*(1), 136–63.

Osborne, M. J. (1995, May). Spatial models of political competition under plurality rule: a survey of some explanations of the number of candidates and the positions they take. *The Canadian Journal of Economics 28*(2), 261–301.

Rabin, M. (1993, December). Incorporating fairness into game theory and economics. *The American Economic Review 83*(5), 1281–1302.

Rabin, M. (1998, March). Psychology and Economics. *Journal of Economic Literature 36*(1), 11–46.

Roth, A. E. (1995). Bargaining experiments. In J. H. Kagel and A. E. Roth (Eds.), *The Handbook of Experimental Economics*. Princeton University Press.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991). Bargaining and market behavior in jerusalem, ljubljana, pittsburgh, and tokyo: An experimental study. *The American Economic Review*, 1068–1095.

Shaw, A. D., J. J. Horton, and D. L. Chen (2011, March). Designing Incentives for Inexpert Human Raters.

In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, New York, NY. USA, pp. 275–284. ACM.

Smith, A. (1761). *The Theory of Moral Sentiments*. A. Millar.

Smith, V. L. and B. J. Wilson (2015). 'sentiments,'conduct, and trust in the laboratory.

Spitzer, M., U. Fischbacher, B. Hermberger, G. Groen, and E. Ferh (2007, October). The neural signature of social norm compliance. *Neuron 56*, 185–196.

Wilson, E. O. (2000). *Sociobiology: The New Synthesis* (Twenty-Fifth Anniversary ed.). Cambridge, Massachusetts, and London, England: Belknap Press of Harvard University Press.

Yenkey, C. B. (2015, March). Distrust and Market Participation. Working Paper, University of Chicago.