In Defense of (Some) Vainglory: The Advantages of Polymorphic Hobbesianism[1]

Gerald Gaus

> "So that in the nature of man, we find three principall causes of quarrell. First, Competition; Secondly, Diffidence; Thirdly, Glory.
>
> "The first, maketh men invade for Gain; the second, for Safety; and the third, for Reputation. The first use Violence, to make themselves Masters of other mens persons, wives, children, and cattell; the second, to defend them; the third, for trifles, as a word, a smile, a different opinion, and any other signe of undervalue, either direct in their Persons, or by reflexion in their Kindred, their Friends, their Nation, their Profession, or their Name."
>
> ~Hobbes, *Leviathan*

1 A JANUS-FACED FEATURE OF SOCIAL COOPERATION

In the *Limits of Liberty* (1975) James Buchanan develops a Hobbesian-inspired analysis of why and how rational individuals would abandon the state of nature and accept a constitutional order. On Buchanan's (1975: 26) reading, Hobbes's analysis assumes that all agents act according to "narrowly defined self-interest" (cf. Chung, 2016). Although Buchanan (1975: 80) insists that his own model does not suppose that each acts only out of self-interest, his guiding aim is to show "how 'law,' 'the rights of property,' 'rules for behavior' might emerge from the nonidealistic self-interested behavior of men" (Buchanan, 1975: 54). Buchanan's insight is that truly rational, self-interested, agents are apt to learn that mutual invasion for gain is a sucker's game; if one can count on others being as sensible as oneself, it is not hard to model an end to the state of war. Narrowly self-interested agents will come to see the possibility and desirability of a Pareto-superior moves from the state of nature: although each prefers more goods to less, and is not concerned with improving the lot of others, each can see that cooperation, not conflict, best promotes one's interests. Sensible, prudent, egoists are the sorts of folks one can do business with and with whom one can reach constitutional terms for ending the state of war. Perhaps they will be tempted to secretly cheat on the rules of peaceful cooperation, but they grasp the critical importance of general compliance with such rules.

Hobbes's state of nature, however, is populated by a second type of agent: glory-seekers who are apt to make war "for trifles, as a word, a smile, a different opinion, and any other signe of undervalue" (Hobbes, 1994: 76). A recent game theoretic analysis of conflict in Hobbes's state of nature identifies glory-seekers as the real root of instability (Chung, 2015). While narrowly self-interested individuals can grasp that conflict leaves them all in a Pareto-inferior position, glory-seekers are willing to turn their backs on

[1] My thanks to Chad Van Schoelandt for comments and suggestions; thanks too to fellow participants at the Workshop on Sharing, University of Manchester and the Workshop on Exploitation, San Diego University. My special thanks to David Wiens for his comments.

mutual benefit, and make everyone worse off, for any "signe of undervalue."

In this essay I argue that vanity is a Janus-faced feature of social cooperation: while, as Hobbes stresses, it certainly can lead to conflict, its very insensitivity to Paretian gains motivates enforcing norms of fairness. A society composed of both egoists and glory-seekers is thus more likely to stabilize fair terms of cooperation than even the most enlightened society of self-interested agents. Rather than, as in many economically-inspired analyses of social order, assuming a society of purely self-interested agents (which, on some views, defines *homo economicus,* see Gaus, 2008: 19-27), we would do better to model polymorphic populations, containing multiple agent types.[2]

Section 2 examines what I call the "Paretian exploitation of egoists." Straightforward egoists of the kind celebrated in accounts of mutual benefit such as Buchanan's are often stuck with accepting very small gains — and we will see why many have thought this is a deep feature of their rationality. However, as is well known, Ultimatum Game experiments indicate that in a wide range of contexts people do not submit to Paretian exploitation: they share, and often in a decidedly egalitarian manner. Section 3 examines several ways that these results have been explained: I suggest that the most satisfying is an account based on social norms of fairness, which enhance cooperation and help self-interested agents avoid Paretian exploitation. This, however, drives us to a deeper puzzle: why do some individuals refuse miserly offers and so uphold fairness norms? Section 4 surveys a number of experiments that have identified negative emotions as critical in the decision to refuse small gains, especially when they run counter to fair sharing. I return to the more general ideas of pride in Section 5, arguing that its critical role in upholding fair share norms is supported by these experiments. I thus advance a hypothesis: an aversion to being undervalued by others — a willingness to turn one's back on schemes of mutual benefit when one feels insulted — is an important support for schemes of fair cooperation, independent of both pro-social egalitarian preferences and to a considerable extent even the normative expectations of others.

2 PARETIAN EXPLOITATION

2.1 Rational Traps

For our purposes, two core commitments of the orthodox conception of rationality are of interest (Gaus, 2011: 63-70).

> *More is Better than Less*: In any given choice Alf will always choose a greater over a lesser value.
>
> *Modularity*: At each point in a decision tree, Alf will choose that course of action which, from that point on, leads to the greatest value.

More is better than less seems basic to the very idea of a rational agent. "The simplest definition of rationality…is that one should choose more rather than less value" (Hardin, 2003: 16). When faced with a choice where the only considerations are between the satisfaction of valued goal G to degree p and the satisfaction of G to level q, where p is greater than q, a rational agent will choose pG rather than qG. *Modularity* is an

[2] Buchanan (1975: 118) sometimes pursues this possibility.

interpretation of *More is Better than Less*: it insists that when a person employs *More is Better than Less*, she is only concerned with, as it were value from "here on out." To see *Modularity* at work — in a case where it seems worrisome to many — consider David Gauthier's (1994: 692) adaptation of a tale from Hume (1976: Book III, Part ii, §5):

> My crops will be ready for harvesting next week, yours a fortnight hence. Each of us will do better if we harvest together than if we harvest alone. You will help me next week if you expect that in return I shall help you in a fortnight. Suppose you do help me. Consider my decision about helping you. I have gained what I wanted – your assistance. Absent other not directly relevant factors, helping you is a pure cost to me. To be sure, if I were to help you I should still be better off than had I harvested alone and not helped you, but I should be better off still if having received your help, I did not return it. This calculation may appear short sighted. What about next year? What about my reputation? If I do not help you, then surely I shall harvest alone in future years, and I shall be shunned by our neighbors. But as it happens I am selling my farm when the harvest is in and retiring to Florida, where I am unlikely to cross paths with anyone from our community.
>
> Being rational persons, we both know this, the scenario I have sketched is one each of us can sketch – and each of us knows it to be true. It would be pointless of me to pretend otherwise. So you know that I would not return your help, and being no sucker, will therefore leave me to harvest my crops alone. Neither of us will assist the other, and so each of us will do worse than need be. We shall fail to gain the potential benefits of cooperation.

The problem can be depicted as in Figure 1.

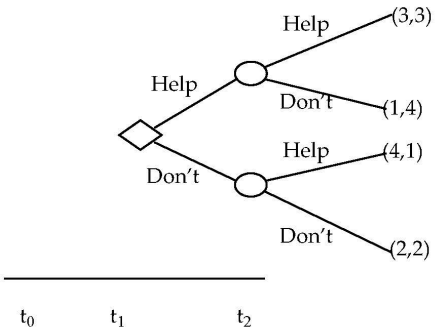


Figure 1: The Hume–Gauthier One-Play Harvesting Game

My neighbor chooses at the diamond, I choose at the ovals; payoffs are ordered from 4 (best) to 1 (worst), first my neighbor's, then mine. The problem is that my neighbor knows the decision tree, and knows that I am modularly rational; once it is my turn to choose I will look to what decision will be best for me from *there on into the future*. If my neighbor helps, I do best by not helping (getting 4 rather than 3). If my neighbor doesn't help, I do best by not helping (getting 2 rather than 1). As in the Prisoners' Dilemma, my dominant strategy is not to help. My neighbor knows this, and so will not help; we are stuck at a Pareto-inferior outcome where neither helps the other.

Gauthier famously argues that *Modularity* should be rejected in favor of the
3) Commitment View, according to which a person can rationally commit himself to a course of action (at time t_0) that, at some point (here t_2), will pursue less value over more. In this case, Gauthier argues, I will get more value by choosing "Help if my neighbor helps." If my neighbor knows at t_0 that I can make such a commitment, she will choose to help, and we will both be better off (3, 3) than if neither helps (2, 2). Unlike in Buchanan's (1975: 136-40) account, the rationality of cooperation does not depend on expected future returns — this is a one-play game with no expectations of future interactions.

Consider now the famous Ultimatum Game, a single-play game between two anonymous subjects, Proposer and Responder, who have X amount of some endowment (say, money) to distribute between them. In a common version Proposer is given an amount of money; he can propose any division he wants. Responder then can either accept or reject. If Responder accepts, both parties get what the Proposer offered; if Responder rejects, no one gets anything. Simplifying, consider (as in Figure 2) an Ultimatum Game in which the Proposer only has two choices about sharing: \$100: (*i*) take 80/offer 20 (*ii*) split 50/50.

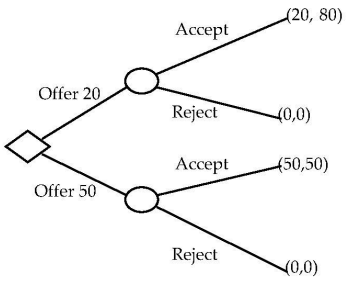


Figure 2: Limited Ultimatum Game

Assuming that the money is the only value under consideration (an assumption that the rest of this paper will interrogate), a modularly rational second player will always choose "accept," since for any offer this is demanded by More is Better than Less; a rational first player will know this, and so she should offer 20, giving her the most (\$80 rather than \$50).

The Harvesting and the Ultimatum games represent different ways in which rational agents can be trapped into unappealing outcomes. In the Harvesting Game they are trapped into a Pareto-inferior outcome; there is a payoff-dominant outcome they cannot reach. In the Ultimatum Game, however, the a rational, narrowly self-interested, second player is trapped into accepting a minimum offer, and a Paretian outcome *is* achieved. It is this "Paretian Exploitation" with which I largely will be concerned. As in familiar cases of exploitation, a bargain is structured in such a way that one party is forced to settle for whatever she can get, no matter how miserly the

offer; it is her own rationality and the structure of the interaction that forces her into accepting the miserly offer.

The Ultimatum Game is not an idiosyncratic case: it sums up a variety of real-world situations in which all can benefit, and share in the fruits of social cooperation, but some are offered take-it-or-leave-it deals by others. Consider, for example, Gauthier's (1986: 190-1) story of the slave society. At one point a member of a slave-owning class, who appreciates the importance of Paretian gains, makes an offer to the slaves: we will stop beating you if you stop trying to escape. The slaves' decision tree would be captured by Figure 2; as modularly rational people they should choose more over less and accept the bargain.[3]

2.2 Play in Ultimatum Games

As is well-known, numerous experiments in diverse settings employing the Ultimatum Game show that Responders very seldom take miserly offers.[4] In the United States and many other countries, one-shot Ultimatum Games result in median offers (of Proposers to Responders) of between 50 percent and 40 percent, with mean offers being 30 percent to 40 percent. Responders refuse offers of less than 20 percent about half the time (Bicchieri 2006: 105). Play in Ultimatum Games does not significantly differ by gender or age; results are strikingly similar whether the stakes are high or low (more on this anon). While those in market societies throughout the world play Ultimatum Games in roughly similar ways, there is much more variance in small-scale, non-market, societies. Indeed, in some small-scale societies (the Machiguenga of the Peruvian Amazon and the Mapuche of southern Chile) the game is played in more "miserly/exploitative" way, as Table 1 indicates.[5]

	UCLA	*Ariz*	*Pitt*	*Hebrew*	*Gadjah*	*Machiguenga*	*Mapuche*
Mean Offer	.48	.44	.45	.36	.44	.26	.34
Modal Offer	.50	.50	.50	.50	.40	.15	.50/.33
Reject Rate	0	—	.22	.33	.19	.048	.065
Reject Offers <20%	0/0	—	0/1	5/7	9/16	1/10	2/12

Table 1: Experimental Results In Ultimatum Games

[3] On Buchanan's (1975: chaps. 2 & 5) account, if the slaves would be enslaved in the state of nature they are rational to accept this offer; if they believe they could successfully rebel and obtain another deal, they have a threat advantage in changing the contract.

[4] Some see this as a major challenge to rational choice theory; see Güth and Tietz (1990). Zamir (2001) objects that investigators rushed to this conclusion, and we have no clear game theoretical prediction as to what fully rational agents would do in ultimatum games.

[5] Data from Henrich and Smith (2004). The Machiguenga and the Mapuche are small-scale societies; the other results are from urban university students in the United States, Israel and Indonesia.

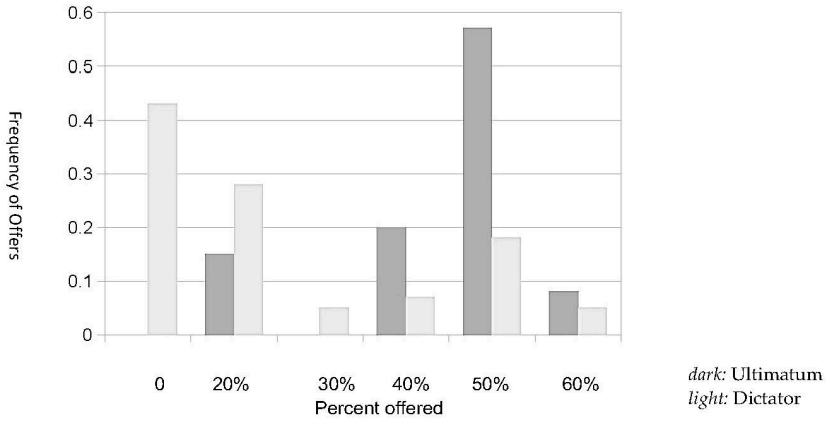
One response to these findings is to see it as evidence supporting Gauthier's Commitment View. A rational Responder can, on the Commitment View, commit ahead of time to rejecting miserly offers, and can rationally carry through on this commitment. If this is generally known, then Proposers would not make miserly offers, knowing that rational Responders will not be trapped by having to make the modular choice for more rather than less. Thus the Commitment View would explain why rational agents are not easily caught in exploitative offers. There are, however, three good reasons to seek to explain these results within the traditional rational choice framework of modular choosers. (*i*) As has been widely recognized, there are a number of problems in explicating the Commitment View as a general theory of rationality; it is one thing to say that it is appealing in special cases, another thing to show just what constitutes a rational commitment, how long a commitment should last, what new information should alter commitments, and so on (Gaus, 2011: 76-86). (*ii*) We may be hesitant about drawing the conclusion that the Machiguenga are less rational than University of Arizona students. They certainly choose differently, but if rationality itself dictates that those who prefer more to less should adopt a Commitment View, then it seems we must attribute some lower level of rationality to the Machiguenga as Responders, or failure to understand the game. Once one builds the solution to these problems into the very concept of rationality, diversity of play becomes, from the point of view of rationality, problematic. (*iii*) Lastly, in one-play anonymous games, when the Proposer does, as sometimes happens (see, e.g., the Hebrew and Gadjah data) make a lower offer, the Commitment View instructs the Responder to choose less (i.e., 0) than what *Modularity* would yield, even though the promise of the Commitment View was that it would yield more to agents than they would receive by following *Modularity*. If one had made a threat to reject and the threat has failed, should one actually make oneself even worse off by following through on the threat?[6] That would seem utterly pointless behavior. Aiming to get more, one gets nothing. In this case the Commitment View seems especially unfortunate. Let us see if we can better explain rational resistance to Paretian Exploitation.

3 EXPLANATIONS FOR MODULAR CHOOSERS

The Commitment View grants that Responders characterized by *More is Better than Less* could care *only* about money, and yet reject low offers. As I have said, one possibility is that the players suffer from defects of rationality or failure to understand the rational strategies of other players (Harrison and McCabe 1996).[7] However, if we are not willing to reject *Modularity* and wish to provide an account of how Responders' rejections could be rational, we should question the assumption that the only value at stake in the game is monetary (Bolton, 1991; van Damme et. al., 2014: 294; Zamir, 2001). Perhaps players have a more complex value scheme in which they prefer more money over less *and* also greater over lesser egalitarian splits. In a classic of evolutionary modeling, Brian Skyrms (1996: chap. 1) showed how a population of players who prefer 50/50 splits could evolve, and how such an evolutionary outcome is more robust than evolutionary

[6] Gauthier (1994) recognizes that threats pose special problems.
[7] For learning in Ultimatum Games, see Eric van Damme et. al (2014: 296ff).

paths that lead to populations in which some are greedy and others take what is left. However, a simple preference for egalitarian outcomes is not well supported by the data. Consider the so-called "Dictator Game" in which Proposer decides on the two shares, and that's the end of the game (not much of a game, to tell the truth). Figure 3 compares typical results in Dictator Games and Ultimatum Games.



Source: Henrich and Henrich, 2007: 166.

Figure 3: Typical Offers in Ultimatum and Dictator Games

In contrast to Ultimatum Games, play in Dictator Games is significantly affected by age and gender. For our purposes, what is important is that when people are guaranteed that their proposal will "be accepted," the modal offer (over .4 of all offers) looks much more like it is determined by straightforward monetary maximization: one takes everything. Yet sharing often occurs (offering 20% of endowment), and a significant number do split 50/50. However, egalitarian sharing is much rarer than in Ultimatum Games, where Responder's choices have to be anticipated by Proposers. This meta-analysis is supported by individual studies, comparing behavior in the two games (Kirchsteiger, 1994). This is not to say that sharing cannot be encouraged in Dictator Games: group norms — and especially whether others in the group are believed to actually share — increases sharing behavior (Bicchieri and Xiao 2009). Moreover, evidence indicates that if affect is primed, and Dictators have less time to think about the decisions, more generous offers occur (Schulz, Fischbacher, Thön and Utikal, 2014).

An important line of inquiry holds that Ultimatum Game egalitarianism is explained by a more complicated valuing of egalitarian outcomes (Fehr, and Schmidt 1999; Fowler, Johnson, and Smirnov 2004). Perhaps people have a general aversion to inequality, but it is much stronger when one gets the short end of the stick. This is a hypothesis with significant support, yet Bicchieri (2005: chap. 3; Bicchieri and Chavez, 2010) persuasively argues that it fails to explain behavior in restricted choice Ultimatum Games. We might contrast two possible hypotheses about why Responders refuse offers: (1) *Equal Outcomes*, according to which Responders prefer roughly equal outcomes and (2) *Norm Violation*, in which Responders are reacting to perceived violation of a norm of fair splits. Both no doubt tell a part of the story but, I believe, overall the data indicate that *Norm Violation* is the fundamental explanation. Consider a modified ultimatum game conducted by Armin Falk et al. in Table 2.[8]

	Proposer's Options		
	Pair 1	*Pair 2*	*Pair 3*
	80/20 50/50	80/20 20/80	80/20 0/100
Responder's Rejection Rate of 80/20 offer	44.4%	27%	9%

Table 2: Rejection Rates Depend on Choices Made by Proposers

In each version of this game the Proposer has only two possible choices. The first in all treatments is to take 80 percent and offer 20 percent; in different versions the paired option is (*i*) a fifty–fifty split, (*ii*) take 20 percent and offer 80 percent, and (*iii*) give everything to the Responder. The Responder knows the Proposer's options. Under pair 1, rejection rates of the 20 percent offer are 44 percent. Note that rejection rate of 20 percent offers drops dramatically when the only option of the Proposers is either to take 80 percent and give 20 percent, or take 20 percent and give 80 percent. If those are the Proposer's options, it does not seem unfair for the Proposer to take the 80 percent for himself, though the inequality of the outcome is the same as under pair 2. And Responders are almost always willing to live with 20 percent given Pair 3, though again the overall outcome is just as inegalitarian as in pair 1. Bicchieri thus concludes that Responders are sensitive to norms: when one gives only 20% when one might have shared equally, one violates a sharing norm, but there is no norm requiring you to sacrifice for the sake of others, in the sense of giving them the lion's share.

It is important that on Bicchieri's account, a social norm is a rule *r* governing some type of behavior in a social network *S*, where most individuals in the social network

[8] Reported by Bicchieri, *The Grammar of Society*, pp. 121–2.

prefer to conform to *r* on the conditions that (*i*) most others in *S* conform to *r* (an empirical expectation) and (*ii*) most people in *S* believe that most others in *S* ought to conform to it (a normative expectation).[9] Condition (*ii*) does not require that anyone in *S* actually believes that others ought to conform to *r* (the definition of a norm does not require that most people hold first order normative beliefs),[10] but that most share a second-order belief about the first-order normative beliefs of others in *S*. Because of this a norm can be based on "pluralistic ignorance" — most people in *S* could have the second-order belief that others in *S* think one ought to conform to *r*, yet it could be the case that no one actually has this first-order belief. The conditions for *r* being a norm would still be satisfied.

The preference to follow *r* is, of course, contextual; it depends on the circumstance for *r*'s application which, we might say, is implicitly a part of *r* (see Cialdini, Kallgren, and Reno, 1990). The preference to follow *r* is a stable part of a person's value function: it is something a person cares about, and which can lead her (as in Ultimatum Games) to forgo monetary benefits in order to follow *r* (say, by rejecting a low offer as a Responder). On Bicchieri's analysis, then, Proposers will tend to give fair offers when they believe that the majority of Responders do, as a matter of fact, reject low offers (the empirical condition) *and* they believe that most others believe that most people normatively disapprove of low offers — the normative condition (Bicchieri and Chavez, 2010). The preference to follow *r* is thus *conditional* upon these two conditions being met. I shall return presently to the importance of expectations.

4 WHY SAY "NO!"?

4.1 A Sense of Justice?

I am a firm supporter of the thesis that we are sensitive to social norms (or, as I tend to say, social rules), and that we tend to punish those who violate them. But the nature of this enforcement mechanism is not well understood. Why are so many individuals in Ultimatum Games so ready to deprive themselves of significant resources in the face of miserly offers, when there is no possibility of compensating gains through future interactions?

An explanation (with some empirical support) that is deeply rooted in political philosophy is that individuals naturally develop a sense of justice — a disposition to comply with, and uphold, just principles and rules (Rawls, 1999: chap. VIII; for empirical support see Carlsmith and Robinson, 2002). We might extend "upholding" to "enforcing" — a person with a sense of justice would go out of her way to approve of action in conformity to norms of fairness and to punish action that violates them. Suppose, then, Responder Betty has a sense of justice: we might expect that if she identifies a certain Proposer, Alf, as one who generally fulfills these social expectations, she will tend to accept Alf's offers, as he is generally a fair-minded person. We can think of her as policing the norm, and so rewarding those who fulfill social expectations. On the other hand, we would expect her, if moved by her sense of justice, to reject the offers a Proposer who has shown himself to disappoint social expectations. If Betty is truly moved by an *impartial* sense of justice, the critical question is not just what offer *she*

[9] This less formal characterization is employed by Bicchieri (2017: chap. 1); for a more formal characterization, see Bicchieri (2006: 11).
[10] Cf. Brennan et al. (2013: 1-14).

receives, but what sort of offers Alf generally makes. If he is a generally fair-minded person, she should still tend to accept a low offer from him — after all, her action is not a response only to his actions against her: his status as a friend or foe of justice is crucial. In an interesting experiment Simon Knight (2012) sought to determine whether Responders were upholding such a sense of justice — whether "the concern is with unfair offers in general" — or were responding to what the Proposer has done to *her* — whether the Proposer gave *her* a high or low offer. Knight found that Responders' behavior supports the latter hypothesis: Responder Betty's action stems from what has been done to *her*, so she will be apt to accept a high offer from a generally unfair Proposer and reject a low one from a generally fair Proposer.

4.2 The Reactive Emotions View

This leads to what we might call the *Reactive Emotions View*: Responders' rejection of low offers is primarily to be explained in terms of Responders' emotional reaction to the offers Proposers make to *them*, in particular whether the offer evokes negative emotions such as anger, irritation, or envy (Bosman, Sonnemans and Zeelenberg, 2001; Kirchsteiger, 1994). General theories of emotion support the anger/irritation/indignation version of this view; as Nico H. Frijda (1996: 311) notes, anger and indignation are generally evoked by norm violation. However, we should distinguish anger from indignation/resentment. Indignation and resentment are distinctly moral emotions that are evoked by norm violation: one can only resent an action if it is perceived as a wrong of some sort, and thus it presupposes a moral evaluation (Strawson 1962). Some see this as a moralized form of anger: we might have anger towards a number of frustrations, impediments, insults and so on, but these need not be moralized.[11]

I have analyzed resentment and indignation at some depth elsewhere (Gaus, 2011: chap. IV); here I shall focus on emotions such as anger, irritation and contempt, which are not inherently moralized. The Reactive Emotions View can be modeled in terms of a two-part value function. Let *X*–*n* be an offer in an Ultimatum Game, where *X* is the total endowment and *n* is the percentage that the Proposer reserves for himself. Then Responder's total value of the *X*–*n* offer will be $\boldsymbol{V}_{MG} - \boldsymbol{V}_{RE}$, where $\boldsymbol{V}_{MG}$ is the value of the absolute *monetary gain*, and $\boldsymbol{V}_{RE}$ is the value based on the *reactive emotions*, a value arising from the negative emotions, which focus on the relation between *X* and *n*.[12] A Responder will accept if total value is positive, reject if it is negative. This supposes that negative emotions are either themselves directly disvalued, or are concomitants of disvalued states (Gaus, 1990: Part I). Thus, for example, an emotional reaction that derives from the Responder's belief that a norm violation has occurred could be the basis of $\boldsymbol{V}_{RE}$;[13] on the other hand, simply seeing the offer as insulting, or getting angry at someone who violates one's expectations in this way would also come under $\boldsymbol{V}_{RE}$.

If we suppose that emotions ($\boldsymbol{V}_{RE}$) are more subject to fluctuation than the value of straightforward monetary resources ($\boldsymbol{V}_{MG}$) — in particular, Responders might "cool

[11] For an experiment focusing on the role of moral anger in trust games, see Thulin and Bicchieri (2016).

[12] We can add positive value that would arise because of pleasure or happiness due to a high offer, treating this as a negative in the second term. As we shall see positive emotions have been measured in Ultimatum-like games, but our real concern is why one would reject an offer where the value of the monetary is above zero, and so what negative (emotional) valuation could drive total value below zero.

[13] The norm regulates the relation between the *X* and *n*.

down" after a period — then we would expect Responders to accept an offer after a cool down period that they would immediately reject. The results of experiments appear contradictory. In an earlier study a break of an hour had no effect (Bosman, Sonnemans and Zeelenberg, 2001) while the more recent study of Veronika Grimm and Friederike Mengel (2011) found a marked decrease in rejection rates after only ten minutes: "While almost no low offers are accepted without delay, a large share (65–75%) of these offers gets accepted after a 10 minutes delay only." Grimm and Mengel also find that low offers of Proposers increase after a break; this is consistent with work on Dictator Games, which indicates that Dictators whose decisions are driven by immediate affect rather than calculation make more generous offers; apparently a cool down period gives each party time to switch into calculation mode, which favors the $\boldsymbol{V}_{\mathrm{MG}}$ element (Schulz et al. 2014). In an experiment on the related "Power-to-Take Game" (see next section) a more complicated pattern emerged: here both a "cooling off" and a "getting steamed up" effect seemed present. If the Proposer's actions are not too miserly from the perspective of the Responder, the Responder seems to cool off after a wait time; however as Proposers get greedier, wait time *raises* the Responders' level of punishment (Galeotti, 2013). If both cooling off and getting steamed up occur, we would expect ambiguous results from wait time experiments.

According to the Reactive Emotions View, low offers, defined as where $X–n$ is (1) a small amount and (2) n is a large proportion of X, should tend to be rejected: $\boldsymbol{V}_{\mathrm{MG}}$ would be low because of (1) and $\boldsymbol{V}_{\mathrm{RE}}$ high because of (2). Conversely, high offers, where $X–n$ is (3) a sizable amount and (4) n is a small percentage of X, should be accepted because $\boldsymbol{V}_{\mathrm{ME}}$ is high (due to 3) and $\boldsymbol{V}_{\mathrm{RE}}$ low (due to 4). This is the generally observed behavior (see e.g., Knight, 2012). But what of offers that are absolutely large, but proportionally low (i.e., in $X–n$, n is a very high percentage of X, but the absolute size of $X–n$ is large)? An important mark against the Reactive Emotions View would seem to be the insensitivity of Responder's behavior in Ultimatum Games to the size of the stakes. One would assume that as $\boldsymbol{V}_{\mathrm{MG}}$ increases (measured, it will be recalled, in absolute size), Responders would be more ready to accept offers, even if n is a high proportion of X. Of course it could be that as the stakes in the game go up so do emotional reactions, but a reasonable hypothesis is that $\boldsymbol{V}_{\mathrm{RE}}$ would not keep increasing as the stakes become higher and higher: one can only get so insulted or angry, but stakes can go up and up.[14] At some point we would expect that $\boldsymbol{V}_{\mathrm{MG}} > \boldsymbol{V}_{\mathrm{RE}}$, and so the (proportionally) "low" offer would be accepted. Yet a variety of studies have shown that play in Ultimatum Games is not very sensitive to the absolute size of the endowments being divided (see e.g., Slonim and Roth 1998).[15] However, as Steffen Andersen et al. (2011) point out, in many of these experiments Proposers advance very few low offers, making it difficult to judge what Responders would do in the face of such offers. In their study, some treatments drastically increased the size of endowments to be divided (equivalent to 1,600 hours of work in India, where the experiment took place) and they elicited many low offers by Proposers. In treatments with traditional sized stakes the behavior of Responders was in line with normal play (though there were more low offers to be rejected); in their very high stakes

[14] For simplicity, I leave aside decreasing marginal utility of money.

[15] This is not to say that stakes have no effect, as stakes rose, "responders (pooled over all rounds) rejected offers less often" (Slonim and Roth 1998: 591), thus supporting a prediction of the Reactive Emotions View.

treatments only 1 of 24 Responders rejected low offers. Figure 4 sums up the predictions of the Reactive Emotions View: offers between x and y should be rejected.

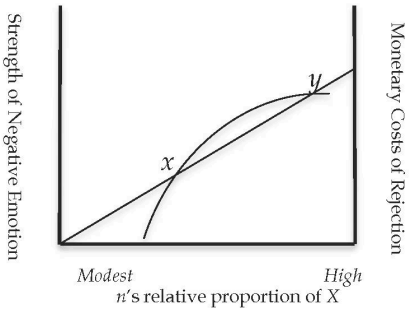


Figure 4: The Reactive Emotions View

4.3 Emotions in Power-to-Take Games

A problem with measuring the role of emotions in Ultimatum Games is that Responders only have a take-it-or-leave-it choice and, as we have seen, low offers are typically uncommon. The role of emotions in Responders' behavior has been extensively studied in a cousin of the Ultimatum Game, the Power-to-Take Game, which allows more scope for emotional reaction. A Power-to-Take Game involves two players, a Taker and a Responder; their roles are determined at random. To start, each player is given an endowment; in some treatments the players earn their endowment in a pre-game task, in others it is simply distributed by the experimenter. Suppose the endowment for each is Y_{Take} and Y_{Resp}. The Taker, then determines take rate — the proportion of the Responder's endowment he will take. The Responder then has an option of destroying any amount of her endowment that she wishes, before the Taker's percentage is transferred from her. So if the endowment was \$10, and the Taker announced a take rate of 50%, the Taker would get \$5 if the Responder destroyed none of her endowment, which would yield total payoffs of \$15 for Taker and \$5 for Responder. If the Responder decides to destroy half her endowment after the Taker announces his take rate, it would reduce her endowment to \$5, of which the Taker would get \$2.50. This game is sometimes described as an Ultimatum Game that allows variable punishment, since Responder can decide on the level at which she will deny Taker's resources.[16] But note that in this game the Responder cannot affect the Taker's endowment, but only the amount of her endowment the Taker can transfer (see Reuben and van Winden, 2010: 908).

In an early pioneering study by Ronald Bosman and Frans van Winden, where

[16] The variability of destruction is meant to uncover the relation of degree of emotional response to degree of punishment; I discuss presently a version of Power-to-Take that gives only limited punishment options which, not too surprisingly, considerably blunts the importance of emotions.

players earned their endowments, out of 39 subjects, only three Takers took 0, positive takings ranged from 25-100%, with a mean of 58.5%, and median 66.7%; 70% was the mode (Bosman and van Winden 2002).[17] Eight Responders chose to destroy part of their endowment, and of these, seven destroyed the entire endowment. In a later study Bosman, Matthias Sutter and van Winden compared this play to another experiment in which endowments were simply distributed at the start of play (Bosman and van Winden 2005). Play in the no effort experiment was markedly different; Takers took an average of 32% more, and many more Responders destroyed, and more opted for intermediate destruction rates. Table 3 summarizes the differences between the effort and no effort experiments.

	Effort	*No Effort*
Destroy Everything	7	6
Destroy Part	1	9
Destroy Nothing	31	25
Total	39	40

Table 3: Results in Two Power-To-Take Experiments (Reuben and Van Winden, 2010).

Especially interesting is that these experiments sought to determine the extent to which emotional reactions explained behavior. Emotions were measured via self-reporting on a seven-point scale ranging from ''no emotion at all'' (1) to ''high intensity of the emotion'' (7). The emotions measured were irritation, anger, contempt, envy, jealousy, sadness, joy, happiness, shame, fear, and surprise (Reuben and van Winden, 2010).[18] The following findings are of interest to us:

- Responders who destroyed report more intense emotional reactions than those who do not.
- The most intense emotions of Responders who destroy in the *no effort* condition

[17] This is typical of takings in Power-to-Take Games; see Reuben and van Winden (2010).

[18] "In both conditions, the sequence of actions was as follows. Before subjects played the one-shot PTT-game, they were randomly divided into two groups. One group was referred to as participants A (the take authorities) and the other as participants B (the responders). Subsequently, random pairs of a responder and a take authority were formed by letting take authorities draw a coded envelope from a box. The envelope contained a form on which the endowment of both participant A and participant B was stated. The take authorities then had to fill in a take rate and put the form back in the envelope again. After the envelopes were collected, we asked the take authorities to report their emotions as well as their expectation of what the responder would do. The envelopes were brought to the matched responders who filled in the part of their endowments to be destroyed. The envelopes containing the forms were then returned to the take authorities for their information. Meanwhile, responders were asked to indicate which take rate they had expected and how intensely they had experienced several emotions after having learned about the take rate. After completing the questionnaires and collecting all envelopes, subjects were privately paid outside the laboratory by the cashier who was not present during the experiment. Experimenters were not able to see what decisions subjects made in the game and how much they earned." (Reuben and van Winden, 2010: 415).

were (in order) anger, contempt, surprise and irritation.
- The most intense emotions of Responders who destroy in the *effort* condition were (in order) irritation, contempt, surprise and anger; the emotions tended to be more intense in this treatment.
- For both treatments, the intensity of these emotions is correlated with the take rate.
- "With effort, the probability of destruction…depends positively on the intensity of irritation and contempt. Without effort, the probability of destruction depends positively on the intensity of anger and contempt, and negatively on the intensity of happiness and joy" (Reuben and van Winden 2010: 420).
- Responders who destroy everything report more irritation than those who destroy only part. Bosman, Sutter and van Winden (2010: 417) indicate that this provides support for what I have called the Reactive Emotions View: this group, they comment, "appear to make a tradeoff between the (emotional) satisfaction of punishment and monetary reward."

In these studies intensity of emotional reactions is a strong predictor of Responder behavior. In a recent study Fabio Galeotti (2015) has shown that the predictive value of emotional reactions can be considerable lessened if the Responders' destroy options are restricted to a fixed rate (2:1) for each unit taken. Rather than Responders deciding how much to destroy in response to a taking, they simply opt to destroy at the fixed rate or not at all. In this treatment negative emotions remain correlated with the take rate, but have less predictive value of punishment. At low levels of punishment (for smaller takings) only contempt was of predictive value; at higher take rates (and so levels of punishment), those with higher levels of anger, irritation and contempt punished more, but this was significantly less predictive than under variable destruction rate treatments. Fixed rate punishment thus appears to blunt the effect of emotions; it especially thwarts Responders' emotionally destroying their entire endowments in response to modest takings.

4.4 Expectations and Fairness

I have suggested that emotional reactions may be an important foundation of behavior to uphold norms. The mere fact that in Power-To-Take Games Responders' destructive behavior is significantly, in some cases powerfully, explained by their emotional reactions does not show that emotions are related to norms. However, data does indicate a connection. Recall the importance of expectations in Bicchieri's account of social norms: a rule *r* is a social norm when the majority in a certain group or social network hold the requisite empirical and normative expectations. Experimental evidence involving Dictator Games indicates that when normative and empirical expectations diverge, there is a strong tendency to align behavior with the empirical expectations (Bicchieri and Xiao, 2009). An important finding in the Power-to-Take Games is that the Responders who punished very strongly tended to be (and in one study were exclusively) those who expected lower take rates than they experienced —recall the presence of surprise (Bosman and van Winden, 2002: 156; Bosman, Sutter and van Winden 2005: 421; Galeotti, 2015: 12). This suggests that while negative

emotions are well correlated with punishing behavior, this is strongly mediated by empirical expectations.

Thus far I have focused on Responders. Reuben and van Winden (2010) studied the effect of Responders' punishment on Takers' take rate in a multi-stage Power-to-Take game. They found that when Responders did not destroy, the Takers who increased their take rate in the second round tended to experience regret after the first round — apparently regretting that they could have taken more and got away with it! Takers who did not experience destruction tended to increase their take rate in the second round. The behavior of Takers who did experience Responder destruction in the first round, however, was complex: some decreased their take rate while others did not. The key appears to be whether the Takers thought their taking was fair or unfair: those who took what they considered to be an unfair amount, to a significant degree reacted to Responders' punishment (i.e., destruction) by decreasing their takings. It is worth pointing out that in the first round these Takers apparently were willing to incur some guilt (say, level Z) in return for high monetary gain X (as they think the offer was unfair, but proceeded anyway, so it would seem ($\boldsymbol{V}_{MG}X > \boldsymbol{V}_{RE}Z$); in the second round they experienced an increased in guilt (Z'), thus it would seem that $\boldsymbol{V}_{RE}Z' > \mathbf{V}_{MG}X$, causing them to lower their taking. However, Responder destruction did not have the effect of lowering the take rate of those Takers who thought their takings fair. This is consistent with other studies concluding that, in addition to the anger of punishers, effective punishment requires violators to experience guilt, say in recognition that they have violated their understanding of fairness or a social norm (Hopfensitz and Reuben 2009). Thus again we are led to the interrelation of emotional reaction and social norms.[19]

5 THE VILE AND CONTEMPTABLE

There is, then, considerable evidence that the emotions of irritation, contempt and anger play an important role in some types of punishing behavior, or, more carefully, in grounding choices that lead one to go away with less (often nothing) rather than accept small gains or allow others to take some of what one possess. Now we might ask, what does this have to do with "vain glory? or, as Rousseau described it "*amour-propre*?"[20] Pride and vanity are not, after all, among the specific emotions studied. But we should not see Hobbes's glory-seeking or Rousseau's *amour-propre* as a specific emotion; it is more of an agent type or personality orientation. Very much in the spirit

[19] Experiments by Thulin and Bicchieri (2016) have shown that "moral outrage" — which is closely related to anger — also seems to underlie third-party compensation behavior, when norm violation has occurred. This is important: we should not suppose that negative emotions must be attached to a preference to punish violators, as opposed to compensating victims. It is important, however, that Thulin and Bicchieri's target emotion appear distinctly moral; in one study emotions were measured, for example, on a 7-pount scale from "Strongly Disagree" to "Strongly Agree" with statements such as "I feel angry when I learn about people suffering from unfairness" and "I think it's shameful when injustice is allowed to occur." These emotions are thus clearly moral emotions, presupposing a normative content.

[20] Amour-propre must not be confused with love of self: for they differ both in themselves and in their effects. Love of self is a natural feeling which leads every animal to look to its own preservation, and which, guided in man by reason and modified by compassion, creates humanity and virtue. Amour-propre is a purely relative and factitious feeling, which arises in the state of society, leads each individual to make more of himself than of any other, causes all the mutual damage men inflict one on another, and is the real source of the "sense of honour." Rousseau (1975: 66).

of Rousseau, William McDougall thought that pride was part of the growth of self-consciousness and a manifestation of the "self-regarding sentiment." As Rousseau might well have said, McDougall (1950: 155) held that "…the idea of self and the self-regarding sentiment are essentially social products; that their development is effected by constant interplay between personalities, between the self and society; that, for this reason, the complex conception of the self thus attained implies constant reference to others and to society in general, and is, in fact, not merely a conception of self, but always of one's self in relation to other selves." This self-regarding sentiment McDougall (1950: 165) maintained, takes two basic forms "which we may distinguish by the names 'pride and 'self-respect'.'' McDougall associated pride with a "positive self-feeling," what Hobbes might call a valuing of the self, which makes one especially sensitive to signs of undervaluing by others and a tendency to insist on one's own way. Pride so construed is high valuing of the self, which is then associated with a tendency to stress a group of specific emotions. Richard S. Lazarus (1991:229) thus observes that "[a]rrogance and smugness, especially the latter, seem to combine with contempt (hence anger) with pride…." Frijda (1994: 89) also notes the association of pride, contempt and scorn. This is not to say that all these emotions are perfectly correlated: in a factor analysis of emotions in Power-to-Takes Games, contempt was the second most unique emotion (after fear), though it still has a .48 and .47 Spearman's rank correlation coefficient with, respectively, irritation and anger (the two emotions it was most closely associated with). However, anger and irritation were themselves much more closely associated (.75) (Galeotti, 2015: 9). Of course if contempt is more pronounced in low takings, this might be expected.

It is perhaps worth noting the importance of contempt in research on Power-to-Take Games. It is an explanatory value in all the experiments we have considered; even in Galeotti's recent study, which minimizes the effect of emotion, contempt remains the sole emotion significantly affecting reactions to small take rates. Recall Hobbes's (1994: 28) claim that contempt sees its object as "vile and Inconsiderable" and the honorable person has "contempt of small difficulties, and dangers" (Hobbes, 1994: 53). To such individuals gains that indicate an undervaluing are vile and inconsiderable, and are to be rejected. "For every man looketh that his companion should value him, at the same rate he sets upon himself" (Hobbes, 1994: 75-6). And when such a person feels undervalued, he is apt to respond destructively, unlike the pure egoist who takes what he can get. "Better nothing than that!" is not a motto of the egoist, and that is why the egoist can get caught in Paretian exploitation and, indeed, submit to takings when he has no choice except submit or engage in self-destructive response.

In the *Limits of Liberty* Buchanan (1975: chap. 8) proposes a solution to this Paretian trap. If the social contract gives an individual especially meager gains over the state of nature, and if the individual has an effective threat to do better by restarting the state of war, she may be able to renegotiate a better deal. But not only is this claim based on highly uncertain calculations, it can lead to further *diminishing* the meager benefits of the social contract: if she would end up enslaved in the state of nature the renegotiation may lead her to make even greater concessions for peace. However, a prideful agent will have contempt, irritation or anger at such vile offers, and so would prefer to destroy her holdings rather than submit. When the prideful are around, hard bargaining can lead to disaster for all. Thus the Janus-faced nature of pride: it can

undergird, as well as undermine, effective social cooperation.

Philosophy
Political Economy & Moral Sciences
University of Arizona

Works Cited

Andersen, Steffen, Seda Ertaç, Uri Gneezy, Moshe Hoffman and John A. List 2011. "Stakes Matter in Ultimatum Games," *The American Economic Review*, vol. 101/7 (December): 3427-3439.

Bicchieri, Cristina 2006. *The Grammar of Society: The Nature and Dynamics of Norms.* Cambridge: Cambridge University Press.

Bicchieri, Cristina 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. New York: Oxford University Press.

Bicchieri, Cristina and Alex Chavez 2010. "Behaving as Expected: Public Information and Fairness Norms, *Journal of Behavioral Decision Making*, vol. 23: 161–178.

Bicchieri, Cristina and Erte Xiao 2009. "Do the Right Thing: But Only if Others Do So," *Journal of Behavioral Decision Making*, vol. 22 (April): 191-208.

Bolton, Gary E. 1991. "A Comparative Model of Bargaining: Theory and Evidence," *American Economic Review*, vol. 81: 109–136.

Bosman, Ronald and Frans van Winden 2002. "Emotional Hazard in a Power-to-Take Experiment," *The Economic Journal*, vol. 112 (January): 147-169.

Bosman, Ronald, Joep Sonnemans and Marcel Zeelenberg 2001. "Emotions, Rejections, and Cooling Off in the Ultimatum Game," at http://hdl.handle.net/11245/1.418488.

Bosman, Ronald, Matthias Sutter and Frans van Winden 2005. "The Impact of Real Effort and Emotions in the Power-To-Take Game," *Journal of Economic Psychology*, vol. 26: 407–429.

Brennan, Geoffrey, Lina Eriksson, Robert E. Goodin and Nicholas Southwood 2013. *Explaining Norms.* Oxford: Oxford University Press.

Buchanan, James M. 1975. *The Limits of Liberty: Between Anarchy and Leviathan*. Chicago: University of Chicago Press.

Carlsmith, Kevin M., John M. Darley, and Paul H. Robinson 2002. "Why Do We Punish?: Deterrence And Just Deserts As Motives For Punishment," *Journal of Personality and Social Psychology* 83, no. 2: 284-99.

Chung, Hun 2015. "Hobbes' State of Nature: A Modern Bayesean Game-Theoretic Analysis," *Journal of the American Philosophical Association*: 485-508.

Chung, Hun 2016. "Psychological Egoism and Hobbes," *Filozofia*, vol. 71: 197-208.

Cialdini, R., C. Kallgren, and R. Reno 1990. "A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior," *Advances in Experimental Social Psychology*, vol. 24: 201–34.

Fehr, Ernst and Klaus M. Schmidt 1999. "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, vol. 114 (August): 817–68.

Fowler, James H., Tim Johnson, and Oleg Smirnov 2004. "Egalitarian Motive and Altruistic Punishment," *Nature,* vol. 433 (6 January): E1.

Frijda, Nico H. 1994. *The Emotions.* Cambridge: Cambridge University Press.

Galeotti, Fabio 2013. "An Experiment on Waiting Time and Punishing Behavior," *Economics Bulletin*, vol. 33/2: 1383-1389.

Galeotti, Fabio 2015. "Do Negative Emotions Explain Punishment in Power-To-Take Game Experiments?" *Journal of Economic Psychology*, vol. 49: 1-14.

Gaus, Gerald 1990. *Value and Justification.* Cambridge: Cambridge University Press.

Gaus, Gerald 2008. *On Philosophy, Politics and Economics.* Belmont, CA: Wadsworth, 2008).

Gaus, Gerald 2011. *The Order of Public Reason.* Cambridge: Cambridge University Press.

Gauthier, David 1986. *Morals by Agreement.* Oxford: Clarendon Press.

Gauthier, David 1994. "Assure and Threaten," *Ethics,* vol. 104 (July): 690–721.

Grimm, Veronika and Friederike Mengel 2011. "Let Me Sleep on It: Delay Reduces Rejection Rates in Ultimatum Games," *Economics Letters*, vol. 111: 113-115.

Güth, Werner and Reinhard Tietz 1990. "Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results," *Journal of Economic Psychology*, vol. 11: 417-449.

Hardin, Russell 2003. *Indeterminacy and Society.* Princeton: Princeton University Press.

Harrison, Glenn W. and Kevin A. McCabe 1996. "Expectations and Fairness in a Simple Bargaining Experiment," *International Journal of Game Theory,* vol. 25: 303-32.

Henrich, Joseph and Natale Henrich 2007. *Why Humans Cooperate.* Oxford: Oxford University Press.

Henrich, Joseph and Natalie Smith 2004. "Comparative Evidence from Machiguenga, Mapuche, and American Populations." In J. Henrich, R. Boyd, S. Bowles, et al. eds., *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societie*s. Oxford: Oxford University Press.

Hobbes, Thomas 1994. *Leviathan*, edited by Edwin Curley. Indianapolis: Hackett.

Hopfensitz, Astrid and Ernesto Reuben 2009. "The Importance of Emotions for the Effectiveness of Social Punishment," *The Economic Journal*, vol. 119 (October): 1534-1559.

Hume, David 1976. *Treatise of Human Nature*, 2nd ed., edited by L. A. Selby-Bigge and L. P. H. Nidditch. Oxford: Clarendon Press.

Kirchsteiger, Georg 1994. "The Role of Envy in Ultimatum Games," *Journal of Economic Behavior and Organization*, vol. 25: 373-389.

Knight, Simon 2012. "Fairness or Anger in Ultimatum Game Rejections?" *Journal of European Psychology Students,* vol. 3: 1-14.

Lazarus, Richard S. 1991. *Emotion and Adaptation.* Oxford: Oxford University Press.

McDougall, William 1950. *Social Psychology,* thirtieth edn. London: Methuen.

Rawls, John 1999. *A Theory of Justice* rev. edn. Cambridge, MA: Harvard University Press.

Reuben, Ernesto and Frans van Winden 2010. "Fairness Perceptions and Prosocial Emotions in the Power to Take," *Journal of Economic Psychology*, vol. 31: 908–922.

Rousseau, Jean-Jacques 1975. *Discourse on the Origin of Inequality*, in *The Social Contract and Discourses*, G.D.H. Cole, trans. London: Dent.

Schulz, Jonathan F., Urs Fischbacher, Christian Thön and Verena Utikal 2014. "Affect and Fairness: Dictator Games under Cognitive Load," *Journal of Economic Psychology*, vol. 41: 77–87.

Skyrms, Brian 1996. *Evolution of the Social Contract.* Cambridge: Cambridge University Press.

Slonim, Robert and Alvin E. Roth 1998. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic," *Econometrica*, vol. 66, No. 3 (May): 569-596.

Strawson, P. F. 1962. "Freedom and Resentment," *Proceedings of the British Academy*, vol. 48: 187–211.

Thulin, Erik W. and Cristina Bicchieri 2016. "I'm So Angry I Could Help You: Moral Outrage as a Driver of Victim Compensation," *Social Philosophy & Policy,* vol. 22 (Spring): 146-160.

van Damme, Eric et. al. 2014. "How Werner Güth's Ultimatum Game Shaped Our Understanding of Social Behavior," *Journal of Economic Behavior & Organization*, vol. 108: 292–318.

Zamir, Shmuel 2001. "Rationality and Emotions in Ultimatum Bargaining," *Annales d'Économie et de Statistique*, no. 61 (January–March): 1–31.