Dignity and the Value of Rejecting Profitable but Insulting Offers

Efthymios Athanasiou
New Economic School
timos.ath@gmail.com

Alex John London*
Carnegie Mellon University
ajlondon@andrew.cmu.edu

Kevin J. S. Zollman
Carnegie Mellon University
kzollman@andrew.cmu.edu

In this paper we distinguish two competing conceptions of dignity, one recognizably Hobbesian and one recognizably Kantian. We provide a formal model of how decision-makers committed to these conceptions of dignity might reason when engaged in an economic transaction that is not inherently insulting, but in which it is possible for the dignity of the agent to be called into question. This is a modified version of the ultimatum game. We then use this model to illustrate ways in which the Kantian evaluative standpoint enjoys a kind of internal stability that the Hobbesian framework lacks. Our interpersonal argument shows that, under certain conditions, Hobbesians prefer to cultivate Kantian commitments in others and promote the presence of Kantians in the population. Our intrapersonal argument shows that agents who are conflicted between Kantian and Hobbesian commitments have powerful reasons not to resolve this commitment in favour of Hobbesian values. Our emulation argument illustrates that in repeated versions of the ultimatum game, the Hobbesian chooses to behave like a Kantian, including publicly repudiating her Hobbesian commitments. Here again, however, the Hobbesian is able to achieve a desired benefit only on the condition that there are genuine Kantians in the population. Finally, our social planning argument explores the reasons why a community of Hobbesians would opt to enshrine a Kantian conception of dignity into law. The paper concludes with some remarks about the policy implications of this work.

> The value or worth of a man is, as for all other things, his price, that is to say, so much as would be given for the use of his power; and therefore is not absolute, but a thing dependent on the need and judgement of another … The public worth of a man, which is the value set on him by the Common-wealth, is that which men commonly call DIGNITY.
>
> Hobbes, *Leviathan X*, 16

> In the realm of ends everything has either a price or a dignity. What has a price is such that something else can also be put in its place as its equivalent; by contrast, that which is elevated above all price, and admits of no equivalent, has a dignity (4:434) … Autonomy is thus the ground of the dignity of the human and of every rational nature (4:436).
>
> Kant, *Groundwork for the Metaphysics of Morals*

[*] Please direct all correspondence on this paper to Alex London at the email address given.

Mind, Vol. 124 . 494 . April 2015
doi:10.1093/mind/fzu186

The concept of dignity features prominently in a broad spectrum of ethical theories and is commonly used to signify how a particular normative framework construes the value or worth of moral agents. As a result, different theories often articulate competing, and in some cases incompatible, accounts of the nature, source, and significance of dignity. These differences are important for a variety of reasons including determining what kinds of behaviours are morally salient and what forms of conduct are permissible or impermissible.

The fact that competing conceptions of dignity are often grounded in different and divergent theoretical frameworks makes it difficult to assess their relative merits without begging key questions. For example, each conception brings with it criteria for ranking different states of affairs as better or worse. It would be question-begging to denigrate the assessments of one theory by appealing to the criteria for making such assessments that are employed by the competing theory. In order to avoid circularity, one might push the argument back a step and consider what can be said on behalf of the respective criteria that each framework employs in evaluating states of affairs. This is to turn the inquiry further in the direction of foundations.

In the discussion that follows we adopt a different strategy. We begin by distinguishing two competing conceptions of dignity. The first view is at home in an ethical tradition that Brian Barry refers to as justice as mutual advantage (see Barry 1989). In this tradition, the dignity or worth of a person, and the regard that they should be shown by others, is ultimately grounded in that person's ability to advance or impede the interests of others. Barry includes in this tradition thinkers such as Hobbes, Hume, Gauthier, Nash, and Braithwaite because they seek, in different ways, to ground either claims to equal regard or claims about the justifiability of unequal treatment, on the roughly equal powers of persons or specific inequalities in bargaining or threat advantage. More contemporary versions of this approach can be found in the work of theorists like Kavka (1986) and Binmore (1989). For convenience, we refer to this as a recognizably Hobbesian conception of dignity. The second view treats the dignity or value of an agent as independent of any advantage that can be gained from social interaction. The notion of human dignity in this tradition grounds duties and constraints that limit the way that inequalities in power, including social or strategic advantages, can be leveraged in social interaction. Because it is widely espoused in the Kantian moral tradition, we refer to this as a recognizably Kantian view of human dignity.

Once these views are on the table, we provide a formal model of a situation in which agents engage in an economic transaction that has the potential to be profitable but also insulting to the dignity of the agent. This is a version of the ultimatum game that has been modified to accommodate decision-makers whose reasoning reflects Hobbesian or Kantian conceptions of their own worth or dignity. The purpose of this mathematical model is to represent clearly the structure of a strategic situation in which an agent's dignity or conception of self-worth might be implicated and to distinguish in precise terms how decision-makers of these different types respond to this situation. We distinguish our account from other strategic situations like the indefinitely repeated Prisoner's Dilemma, social contract bargaining, and commitment problems.

We use this model to illustrate ways in which the Kantian evaluative standpoint enjoys a kind of internal stability that the Hobbesian framework lacks. In static, single-shot instances of the ultimatum game the Kantian is able to secure a larger share because she will reject offers that are insulting. The Hobbesian respondent would like to reap similar benefits. Our *interpersonal* argument shows that the Hobbesian respondent can do this only if there are enough real Kantians in the community and therefore that, under certain conditions, the Hobbesian not only prefers the company of Kantians, but prefers to cultivate Kantian commitments in others. Our *intrapersonal* argument shows that agents who are conflicted between Kantian and Hobbesian commitments have powerful reasons *not* to resolve this commitment in favour of Hobbesian values. Our *emulation* argument illustrates that in repeated versions of the ultimatum game, the Hobbesian chooses to behave like a Kantian, including publicly repudiating her Hobbesian commitments. Here again, however, the Hobbesian is able to alter her behaviour only on the condition that there are other real Kantians in the population. Finally, our *social planning* argument explores the reasons why a community of Hobbesians would opt to enshrine a Kantian conception of dignity into law.

In order to illustrate the practical implications of this highly idealized example, we close with some reflections on a claim recently defended by Alan Wertheimer. Wertheimer's claim is that there should be a presumption against interfering in mutually advantageous transactions that are freely undertaken by informed and consenting parties, even if those transactions are exploitative, unfair, or unjust (see Wertheimer 2008, p. 84; Wertheimer 2011, pp. 214–23). The model that we present here illustrates one mechanism through which social policies that prohibit beneficial but insulting offers might work to the advantage of individuals who might be likely targets of such offers and to the advantage of the communities in which such agents reside.

The arguments that we present are unique in several respects. Our model treats the Kantian and the Hobbesian as rational in the same sense and allows us to focus on the way their competing conceptions of dignity influence their choices. This allows us to illustrate why, in the single-shot version of the ultimatum game, Hobbesians who might seek to adopt a Kantian conception of dignity for Hobbesian reasons would fail to act on those values. Our arguments differ, therefore, from traditional two-tiered approaches in which it might be argued that the Hobbesian has Hobbesian reasons to adopt Kantian values. Rather, our arguments show that in some situations, Hobbesians have Hobbesian reasons to repudiate their own values and to promote genuine Kantianism in others. This argumentative strategy may not persuade fully committed Hobbesians to reject their preferred conception of dignity, but it provides agents who have not made up their minds about the relative merits of these different theoretical frameworks with compelling reasons not to fully commit to the Hobbesian point of view. Finally, the methodology that we use in this paper could be extended in interesting ways. In particular, our results might provide a starting point for those who are interested in naturalistic accounts of ethical theories, or evolutionary accounts of how Kantian and Hobbesian moral standpoints might emerge within a particular environment.

1. Two concepts of dignity

In this section we describe the two models of human dignity that are the focus of this paper. In particular, we want to convey the sense in which one is recognizably Kantian and the other recognizably Hobbesian. We say 'recognizably' because we are not making the more ambitious claim of providing completely adequate models of the concept of dignity within Kantian and Hobbesian ethics, since, as we noted above, the concept of dignity lies at the intersection of many different profound and controversial features of these competing and often conflicting theories of value. Rather, the models that we offer are simplified accounts of core features of each of these theories and our claim is that they are sufficiently representative of these views that, within the choice situation on which we focus, they enable us to make an argument of philosophical importance.

1.1 The Kantian conception of dignity

The concept of dignity plays a central, if not a defining role in Kantian ethics (see Hill 1992, pp. 10–11; Cummiskey 1996). For Kant, the realm of value can be mapped with a two-fold distinction (see Korsgaard 1996, pp. 249–74). First, we can distinguish things that are valuable as ends and things that are valuable for their relationship or contribution to such ends. Second, among the things that are valuable as ends, we can distinguish those that have a price and those that have a dignity. Things that have a price are fungible; they can be replaced by something of equivalent value in two senses. First, these ends can usually be replaced, without loss, by a different token of the same type. There is nothing uniquely valuable about any particular pair of new size 8 Nike sneakers that could not be fully replaced by a different pair of new size 8 Nike sneakers. Second, it is possible to represent the value of such things in terms of an equivalent amount of some other thing. In this case, for example, each pair of sneakers has the same value as a certain amount of money.

In contrast, the Kantian holds that ends that have dignity are uniquely valuable and that their value is above all price (see Hill 1992, pp. 204–6; Hill 2003, pp. 24–7, 42–3, 220–3). For moral agents, their rational nature represents the source of the value of things or goods. This status as beings that set their own ends gives rational agents a value that cannot be replaced or substituted without loss by another token of the same type. Moreover, the Kantian holds that each entity with a dignity value is above all price in the sense that the value of such an agent cannot be given an equivalent in terms of some amount of things with a price value. The key claim here is not that the value of things with a dignity cannot be compared to the value of things with a price. It is, rather, that they *can* be compared and that things with a dignity value are categorically or lexically more valuable than things with a price.

This distinction in value plays a fundamental role in Kantian ethics. Respect for the dignity of agents is seen as grounding numerous constraints on the behaviour of agents, both in terms of their conduct toward themselves and others (see Darwall 2006, esp. p. 292). In particular, the core of Kantian morality consists in showing proper respect for the distinct value of moral agents, where this means treating them in ways that show their categorically superior value to any practical end that we embrace as a means of advancing our welfare or happiness. The imperative to respect the status or value of agents as ends in themselves grounds a range of duties to the self as well as duties to others. In the case of others, we may not bring about ends that we value through the use of force, fraud, coercion, or manipulation. In the case of the self, the Kantian claims that there are duties not to engage in behaviour that debases, degrades, or devalues one’s own rational nature. In both the intrapersonal and interpersonal cases, the wrongness of these actions is explained by the claim that such conduct treats agents as though they were sophisticated tools with a use value that could be matched or outweighed by the value of the ends the agent is seeking to advance.

For our present purposes, we model the Kantian as recognizing two dimensions of value, a dimension for things that have a price and a dimension that is related to dignity. We then focus on a particular choice situation in which two parties must divide a good that falls squarely within the dimension of price. The idea is to avoid focusing on a transaction or interaction that the Kantian views as impermissible per se. So we do not focus on cases in which agents sell their sexual services, or their body parts, or other forms of labour that the Kantian might regard as debasing or degrading. Rather, we focus on a transaction that the Kantian does not object to per se, but in which the Kantian’s concern for her own dignity might become salient and lead her to make a choice that would distinguish her from the Hobbesian.

The choice situation that we focus on is known as the ‘ultimatum game.’ In this game,[1] two parties are given an amount of money (call it a ‘dollar,’ but this could be proxy for many dollars) to divide among themselves. If they can agree on a division they can keep the share that each receives. If they cannot agree, nobody gets anything. In the ultimatum game, however, one agent (the proposer) is given the advantage of proposing a division of the dollar to the other player (the respondent) who then has the ability to either accept the offer — and walk away with the corresponding profit — or reject it — and forgo the proposed profit while denying any profit to the proposer. In all the examples we examine, we focus on the behaviour of the agent, either the Kantian or the Hobbesian, when he or she is in the position of the respondent.

[1] The ultimatum game has come to prominence in the field of experimental economics following the work of Guth et al. (1982). Thaler (1988) discusses the challenges that the ultimatum game presents for economic theory.

The ultimatum game models a wide variety of common exchange situations. For instance, if two people come to learn that 'the seller' would be willing to sell an item for $5 (or more) and that 'the buyer' would be willing to buy that same item for $10 (or less), both have come upon a surplus. If the seller were to sell the item to the buyer for $7 the seller would feel as though she had gained $2 (since she would have been happy to sell for $5) and the buyer would feel as though he had gained $3 (since he would have been willing to pay $10). If the seller has an item that the buyer could not secure anywhere else, and she is able to set a take-it-or-leave-it price, then the buyer and seller are playing an ultimatum game. A wily seller who cares only for money and who knows the buyer's maximum price will realize that she can price the item at $9.99 and reap the lion's share of the surplus generated by the situation.

The ultimatum game is of significant interest to a wide range of scholars because people who are placed in the position of the recipient routinely reject divisions as high as 70/30 (see Murnigham 1982). Behavioural economists, social psychologists, and others are at great pains to understand why it is that recipients routinely turn down profitable but highly unequal offers. We do not use the ultimatum game to explain this empirical phenomenon. Rather, we use it to model an empirically plausible situation in which the difference between the two conceptions of dignity discussed here are manifest in choice and conduct.

In particular, we model the Kantian as willing to engage in an economic exchange in this context and as willing to accept an unequal division of the surplus. So even when the proposer presses his advantage and tries to secure a larger division of the money for himself, we model the Kantian as willing to see the transaction as simply a matter of profit and loss. However, in our model there is a point where the division is so unequal that the Kantian believes that she is being treated with disrespect and that to accept the offer would be to debase herself in some way. In the formalism we present below, we label this point r. Although r is an amount of money, it serves as a signal to the Kantian that the interaction has shifted from a purely economic exchange, to one in which the proposer is acting in a way that fails to respect the recipient as a moral equal.

Again, we are not claiming that this model explains why some people actually reject profitable but highly unequal divisions in the ultimatum game, although that empirical hypothesis might be worth exploring in the future. The only potentially controversial claim to which we are committed is that there are economic transactions that the Kantian does not view as an affront to dignity per se, but which can be conducted in such a way that the Kantian comes to see them as implicating her dignity, and that in such a case the Kantian would be willing to reject a profitable but insulting offer.[2]

To illustrate the plausibility of this claim, consider the following example. Suppose a plumber normally charges $50 an hour for his services. In special situations he will donate his labour for free to help someone in need, and in order to get lucrative contracts he will even accept a lower wage than he thinks is fair. Today, he has no appointments and has made no plans for the day. He will likely waste the day watching television. The plumber does not enjoy watching television, he would prefer to work, but if there is no work it is how he will pass the time. His wealthy neighbour knows the plumber has no appointments for the day, and so offers to pay the plumber $6 an hour to move a sink in his bathroom. The plumber explains that this is a six-hour job and would require a significant amount of labour. He offers to work for less than his regular wage, but not so low as $6. The plumber decides privately that he would accept as little as $25 an hour. The neighbour explains that he knows the plumber has no other opportunities today, and that he knows the plumber dislikes being idle. The neighbour reiterates his $6 an hour offer, declaring ‘take it or

[2] Strictly speaking we require a slightly stronger claim, namely, that the Kantian must also reject those offers which implicate her dignity. One might object that in special cases a Kantian might accept an offer that insults her dignity without degrading herself. This might be the case, for example, in situations where the Kantian needs the benefits on offer in order to survive or to meet an important need (we thank an anonymous referee for raising this challenge).

Three replies to this objection are in order. First, the model that we outline here can accommodate the claim that in dire-enough circumstances Kantians may sometimes accept insulting offers. This simply requires that the α parameter that we introduce in Sect. 3 be interpreted (in the interpersonal argument) as the probability of interacting with a Kantian who is not in a sufficiently dire situation. The interpretation of α in our intrapersonal argument already reflects the fact that a conflicted agent sometimes chooses as a Hobbesian.

Second, however, it is not clear to us that the Kantian ought to accept such offers. If the offer amounts to treatment that does not reflect the recipient’s status as a member of the kingdom of ends, then the Kantian should not be complicit with such treatment. The fact that this may redound to the detriment of the agent’s welfare simply reflects the extent to which the Kantian moral standpoint differs from a welfarist consequentialist standpoint.

Third, there may be a legitimate dispute among Kantians over this point, and how this dispute plays out may matter to those who have not yet committed to one of these conceptions of dignity. In this case, the arguments that we provide might be viewed as lending support to the value of adopting a stricter Kantian attitude toward affronts to one’s dignity. That is, the Kantian preserves her dignity, which is centrally important, and also secures greater monetary or material benefits — something the Kantian also values.

leave it.’ The plumber believes that to accept such a low wage would be demeaning. He refuses the offer and instead watches television all day.

In this case the plumber is the recipient in the ultimatum game and he rejects a profitable offer in favour of an outcome in which he simply wastes the day. In this case the plumber’s explanation of his behaviour was that his neighbour’s offer treated the plumber’s labour as worth far too little and was so small as to impinge the plumber’s dignity.

It is also important to note that we are claiming only that the Kantian will reject certain divisions of the dollar that she judges to be insulting or disrespectful. We are not making the stronger claim that the Kantian can accept only ‘fair’ divisions. While this might be a plausible claim to make from the Kantian standpoint, we want to sidestep contentious questions about what constitutes a fair division and to remain agnostic about whether there can be offers that are unfair but not insulting.

1.2 The Hobbesian conception of dignity

In contrast to the Kantian, the person we call the Hobbesian takes the view that the standard for valuing agents is the same as the standard for all valuation, namely, how will various ways of interacting with an agent advance or frustrate the decision-maker’s own ends. The quote from Hobbes with which this paper opens gives succinct expression to the sense in which there is no stark, categorical difference in the value of agents and the value of other things. The value or worth of an agent is similar to the value of a complex tool; it is a function of the degree to which that agent is needed by, relied on, or is capable of advancing or frustrating the goals, ends, or interests of others.

The reasons the Hobbesian recognizes to curb or constrain her conduct, either with respect to its effects on herself or others, do not emanate from a conception of respect for agents as a unique kind of thing with a distinct sort of value. This is powerfully illustrated by a passage from Hume in which he describes a ‘species of creatures intermingled with men, which, though rational, were possessed of such inferior strength, both of body and mind, that they were incapable of all resistance, and could never, upon the highest provocation, make us feel the effects of their resentment’ (see Hume 1739, p. 190). For Kantians, the rational nature of such beings would make a claim on others of sufficient force to constrain the way it is acceptable to treat them. But for Hume, the fact of such radical inequality in power or ability translates into a radical inequality in standing. With respect to such creatures we would not be bound by considerations of justice and they could possess no rights or property because ‘our intercourse with them could not be called society, which supposes a degree of equality.’ For both Hume and Hobbes, the relevant space of equality is the space of power or ability, and the radical inability of these creatures to adversely affect the interests of those who might abuse them relegates them to an utterly subordinate status.

In the ultimatum game, we model the Hobbesian as simply trying to maximize her monetary profit. The Hobbesian recognizes that the person in the position of the proposer has been gifted a tremendous strategic advantage and that the respondent is in a very weak position. As a result, in the single-shot ultimatum game, the Hobbesian respondent will accept any positive offer. The reason is simply that the Hobbesian recognizes that in this situation she lacks any leverage to improve the size of the division that she receives and that any positive offer represents a gain over the default position of walking away with nothing.

2. The economic rationality of the agents

For the purposes of the present inquiry, we model both the Hobbesian and the Kantian as being rational in the same sense. This point is worth clarification since the nature of rationality and rational choice is likely to be a subject over which the Kantian and the Hobbesian disagree. In particular, the Hobbesian may be inclined to dismiss the Kantian out of hand on grounds of irrationality. For example, assume that the Hobbesian observes the Kantian in two different ultimatum games. In the first game, the Kantian and her partner must divide a nickel and the Kantian accepts a proposed division in which she receives 2 cents and the other party 3. In the second game, the pot is $100 and the Kantian rejects an offer of 2 cents. Because the Hobbesian takes the Kantian’s actions in the first game to express the Kantian’s preference for 2 cents over nothing and the second game to express a preference for nothing over 2 cents, the Hobbesian may simply regard the Kantian as irrational.

Characterized in this simple way, the Kantian is violating a rather basic canon of rationality. She cannot rationally prefer x to y and simultaneously prefer y to x. But as we model the Kantian, this is only an apparent irrationality. In the formalism we introduce below, these offers differ in their relation to the point r at which the Kantian sees the proposer’s behaviour as implicating her dignity. In the one case, the offer of 2 cents was not insulting because it represented almost half of the amount available to be split. To our Kantian, this split did not evoke concerns about whether her status as the moral equal of the proposer was being called into question. In the second case, to accept the offer of 2 cents would be undignified because the relative gain is so disproportionate that our Kantian viewed it as disrespectful.

We attempt to capture the preferences of the Kantian by treating every outcome of the game as being comprised of two parts: a dignity part and a pecuniary part. The dignity part is binary: either the outcome is insulting or it is not. The pecuniary part simply represents the absolute dollar amount offered. The Hobbesian prefers those outcomes that yield higher monetary values (or higher expected monetary values in the case of gambles) regardless of whether or not those offers would be considered insulting by the Kantian. The Kantian, on the other hand, prefers any outcome that preserves her dignity to any that is insulting regardless of amount. When offers are not insulting, that is, when they are above *r*, the Kantian does not perceive her dignity to be at issue and thus prefers more money to less, like the Hobbesian. When offers are insulting, that is, when they fall below *r*, the Kantian faces a choice between monetary gain and a loss of dignity, on the one hand, and foregoing monetary gain while preserving her dignity, on the other. In our model, the Kantian in this situation always prefers to retain her dignity and to forgo any monetary gain.

It might be objected that our approach is unfair to the Kantian, because it appears from the Kantian's behaviour that she ascribes a certain monetary value to her dignity, namely *r*. Although the Kantian's *behaviour* is equivalent to someone who thinks that her dignity is worth *r* dollars, our model does not presume or require that the Kantian makes such an assignment. That is, we treat *r* as the point at which the Kantian shifts her attitude toward the proposer from one in which only monetary stakes are at issue, to one in which her dignity is at issue. This does not necessitate that the Kantian has a monetary valuation for her dignity. The Kantian's dignity valuations and her monetary valuations are non-continuous in the economic sense of that term.[3]

[3] The preferences of both the Kantian and the Hobbesian are each represented by a relation $\succsim_i$ that is *complete* and *transitive*. Moreover, for both, other things being equal, 'more money is better than less,' that is $\succsim$ is *monotonic* with respect to money. Formally, for each

Both the Hobbesian and the Kantian have complete and transitive preferences over all outcomes in the ultimatum game, which are defined as dignity–dollar pairs. This is consistent with the most basic economic sense of rationality. But many economists demand more than complete and transitive preferences over certain outcomes — they also want to know how the Hobbesian and the Kantian value gambles over outcomes. For our Hobbesian this is no problem; we assume that he will maximize his expected monetary gain. For our Kantian things are more complex. How Kantians should value gambles in which different ranges of benefits can be secured only at the prospect of suffering an insult to one's dignity with differing probability remains an open question.

This is a thorny issue. Should the Kantian be willing to allow some small probability of a loss of dignity in exchange for some positive monetary gain, she might be accused of placing a particular price on her dignity — something which many Kantians wish to avoid. On the other hand should the Kantian be unwilling to tolerate any chance of dignity loss, no matter how small, for any monetary gain, no matter how large, the Kantian will violate some of the basic axioms of rationality commonly used in decision theory.

In the analysis that follows, we sidestep this issue. Nothing in our analysis requires that we take a position on how the Kantian would choose in the face of mixtures of options. The choice behaviour that we rely on in the argument below is consistent both with a Kantian who follows the standard decision-theoretic axioms and with theories of rationality that are weaker than the standard account (see Sen 2002 and Levi 1986). We therefore take it as an advantage of our analysis that it deals with a decision context in which broader and more controversial questions about the nature of rationality can be bracketed, so that Kantian and Hobbesian views of dignity can be evaluated in their own right.

$i \in \{\text{Kantian}, \text{Hobbesian}\}$ and each $(x', d), (x'', d) \in \mathbb{R} \times \{0, 1\}$

$$(x', d) \succsim_i (x'', d) \text{ if and only if } x' \geq x'', \text{for each } d \in \{0, 1\}$$

However, for the Kantian and for each $x', x'' \in \mathbb{R}$, $(x', 1) \succ_K (x'', 0)$. Put plainly, there exists no amount of money that would make the Kantian willing to suffer a blow to her dignity. An economist would say that $\succsim_K$ is non-continuous and, in particular, *lexicographic*. For the Hobbesian instead, for each $x' \in \mathbb{R}$ there exists some finite $x'' \geq x'$ such that $(x', 1) \sim_H (x'', 0)$. Put plainly, the Hobbesian attaches a monetary value to her dignity.

3. A static generalization of the ultimatum game

The traditional economic analysis of the ultimatum game usually uses the Hobbesian as a benchmark. If a Hobbesian proposer[4] knows she is facing a Hobbesian, she knows that the respondent prefers more money to less. From this the proposer can conclude that the Hobbesian respondent will accept any offer that leaves him with positive return and even potentially an offer which leaves him with nothing (since the Hobbesian respondent is indifferent between accepting and rejecting in such a case). As a result, the proposer will offer the smallest positive amount possible, or alternatively propose giving nothing to the respondent if he would accept. This reasoning pattern (known as backward induction) picks out a unique equilibrium of the game.[5] This clearly leaves the respondent with relatively little.

If a Hobbesian proposer knows that the respondent is also a Hobbesian the proposer will keep almost all of the good. The Hobbesian respondent will regret that outcome, in that he would have preferred more to less, but he will make no moral judgement about the propriety of the proposal — the proposer was simply taking advantage of the superior bargaining position afforded her.

We will now turn to the situation where the respondent is a Kantian. For the sake of concreteness we suppose in this game that there is a threshold, r, such that any proposal which leaves the Kantian with less than r is regarded as insulting. If the Kantian receives an offer above r, then she gets a monetary gain and she keeps her dignity intact. If she accepted an offer below r, however, she would lose something that she regards as extremely important — so important that she would rather forgo a monetary gain in favour of retaining her dignity. Intuitively, then, when the Kantian rejects an offer that falls below r she is not simply incurring a monetary loss. She is preserving her

[4] We will henceforth assume that the proposer is a Hobbesian without explicitly mentioning this assumption. We are primarily focusing on an argument in favour of a Kantian respondent, and this argument is made most difficult by assuming the proposer is Hobbesian. In particular, if the proposer is a Kantian, and Kantians will not propose disrespectful divisions, then the Hobbesian would benefit more directly from the presence of Kantians in the community. Our argument is still valid if the proposer could either be Hobbesian or Kantian.

[5] We will throughout assume that if the respondent is indifferent between accepting and rejecting then he will accept. This is merely a simplifying assumption which makes stating the results significantly simpler. Allowing an indifferent respondent the option of rejecting would make no significant difference in our conclusions so long as we allow that money is only finitely divisible. The equilibrium we describe is the unique sub-game perfect equilibrium, although other perfection criteria will select it as well.

dignity. There are various ways that this intuitive difference can be represented mathematically. The Kantian could be represented as using a vector of distinct utility functions, one for money and another for dignity. The Kantian could also be represented as having a single utility function for money and dignity and as incurring a significant loss in utility if she accepts an offer below *r*. We have adopted the latter approach only because it is computationally more tractable to work with a single utility function.[6]

In addition, rather than work with negative utilities, we have rescaled the utility function so that accepting an insulting offer, rather than being negative utility, is now the 0. This means that rejecting an insulting offer leaves the Kantian with a positive utility that is represented by *r*. That positive value would also be present in any offer that is above *r*. Rescaling in this way is a purely mathematical convenience that captures the intuitive idea that by rejecting an insulting offer, the Kantian retains something of significant value.

If a Hobbesian proposer was now aware that she was facing a Kantian respondent, she would know that any offer below the threshold *r* would be rejected — the Kantian would prefer to preserve her dignity rather than accept an insulting offer. However, an offer of *r* or above will be accepted, and so the proposer will offer *r* to the respondent. Because this (just barely) does not insult the Kantian respondent, she will accept the offer because it leaves her with some positive financial gain.

Now we turn to a yet more complex possibility. Perhaps the proposer is uncertain as to whether the respondent is a Kantian or Hobbesian. We model this with the game depicted in Figure 1. Nature moves first and determines whether the recipient of the offer is a Kantian or Hobbesian. He is a Kantian with probability α. Although the proposer is not informed of which type she is facing, the value of α is common knowledge. We assume that $\frac{1}{2} \geq \alpha > 0$. The proposer moves next. Her proposal consists of a number x in the closed real line interval $[0, 1]$, the share of a dollar she proposes to

[6] We do not mathematically model the Kantian as considering the two dimensions of the offer explicitly discussed before, but our Kantian behaves exactly as if she does and thus nothing is lost *from the perspective of our model*. If we were to undertake the mathematically more difficult process of representing the Kantian's reasoning explicitly, we would be able to reproduce exactly the same argument. The only additional assumptions we require in order to use lexicographic preferences is that the set of non-insulting offers is closed (an offer of exactly *r* is regarded as non-insulting). This is a technical assumption that arises from idealizing the split as having infinite possible divisions. Again, this assumption does not drive any of the results, it merely simplifies the mathematics.

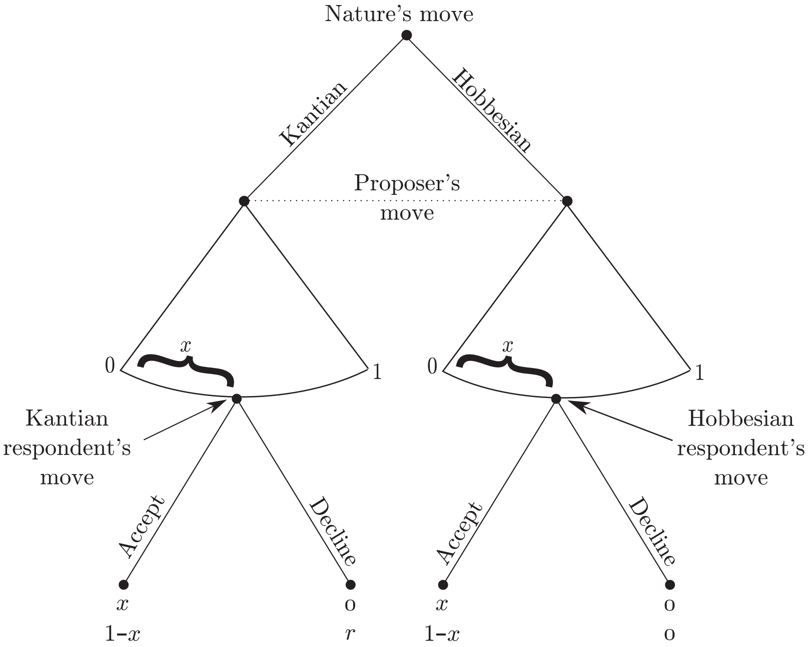

Figure 1: A generalization of the ultimatum game. The first move is a move by Nature to determine whether the responder is a Kantian or Hobbesian. The proposer moves second, in ignorance of Nature's move. The proposer chooses a value x. The respondent, who knows her type and the proposal, chooses whether to accept or decline the offer. If the respondent accepts, the proposer receives x and the respondent receives $(1 - x)$. If the respondent declines the proposer receives 0 and the Kantian or Hobbesian respondent receive r or 0 respectively

retain for herself. The recipient of the offer may choose to accept, in which case his payoff is $1 - x$, or decline. In that latter case the dollar is burned, which leaves the proposer with zero, while the recipient, depending on whether he is a Kantian or Hobbesian, obtains zero or r respectively. We assume that r is in the open interval (0, 1). If either r or α were equal to zero, the set up would correspond to the classic ultimatum game.

Our basic interpretation of Nature's move is that the proposer is interacting with a person who represents a random member of a population that contains both Kantian and Hobbesian types (represented as R_K and R_H respectively). Both parties know ahead of time what the probability is of a proposer interacting with a Kantian or Hobbesian respondent, but the proposer does not know exactly which type she is currently facing. Both parties also know the value of r, the point at which the offer would be insulting to a Kantian respondent. This may be an idealizing assumption, but relaxing it would not have a radical effect on the underlying results of the model.

The equilibrium of the game is described in Proposition 1. It depends on whether $\alpha < r$ or $\alpha \geq r$. However, within each case it is unique. If $\alpha \geq r$, the equilibrium outcome is such that the proposer concedes r to the recipient. In this case, the proposer finds that the prospective cost of having an insulting offer rejected outweighs the benefit of fully exploiting his negotiation power. This small victory on dignity's behalf does not depend on the result of Nature's move and, consequently, both types of recipient reap the rewards. That is to say, the Hobbesian finds himself obtaining r dollars, although he would have accepted an offer involving zero dollars.

Proposition 1: Assume that the recipient of either type, if indifferent, opts for yes. Moreover, assume that the proposer, if indifferent between any two actions x^1, x^2, with $x^2 > x^1$, opts for x^1. Under these assumptions:

(1) if $\alpha < r$ the game has a unique Bayes-Nash equilibrium involving the following strategies:

R_H : 'yes' for each $x \in [0, 1]$
R_K : 'yes' if $x \in [0, 1 - r]$, 'no' otherwise
$P : x = 1$

(2) if $\alpha \geq r$ the game has a unique Bayes-Nash equilibrium involving the following strategies:

R_H : 'yes' for each $x \in [0, 1]$
R_K : 'yes' if $x \in [0, 1 - r]$, 'no' otherwise
$P : x = 1 - r$

Proof:
The Hobbesian recipient's best response is to accept any offer. The Kantian recipient accepts if an offer is such that $1 - x \geq r$ (equivalently if $1 - r \geq x$) and declines otherwise. Therefore, we may determine the proposer's expected payoff. In Figure 2 we discriminate between two cases: $r > \alpha$ (left) and $\alpha > r$ (right). The bold curve depicts the expected payoff of the proposer as a function

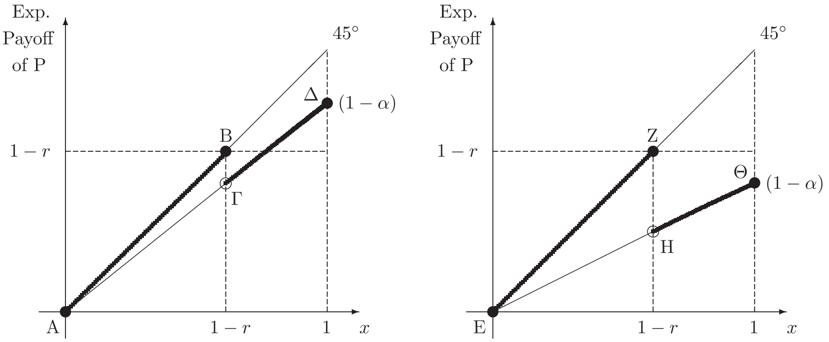

Figure 2: The expected payoff of the Proposer as a function of x is the depicted by the bold curve for two values of the parameter α. On the left hand side $\alpha < r$ and therefore $(1 - \alpha) < (1 - r)$. On the right hand side $\alpha > r$ and therefore $(1 - \alpha) > (1 - r)$. Points on the line segment (ΓΔ) are equal to $(1 - \alpha)x$, and points on the line segment (HΘ) are equal to $(1 - \alpha)x$

of x. The coincidence with the 45 degree line reflects the fact that any proposal such that $x \leq 1 - r$ is accepted with certainty, or, in other words, independently of the recipient's type. On the contrary, an offer such that $1 - x < r$ is accepted with probability $1 - \alpha$, resulting in an expected payoff equal to $\alpha \times 0 + (1 - \alpha) \times x$. If $\alpha < r$ the expected pay-off of the proposer is represented by the union of the bold lines (AB) and (ΓΔ), excluding, in line with our assumption, point Γ. If $\alpha > r$ the expected pay-off of the proposer is represented by the union of the bold lines (EZ) and (HΘ), excluding point H. One can verify with the help of Figure 2 that the proposer maximizes his expected payoff at point Δ if $r > \alpha$, or at point Z if $\alpha > r$. Both these points represent the unique maximum. If it is the case that $\alpha = r$, the proposer maximizes by setting x equal to either 1, or $1 - r$, or mixing between the two. All these options are pay-off equivalent. By assumption, the proposer chooses $x = 1 - r$. ■

3.1 The interpersonal argument

Since the proposer is unaware of the type of the respondent, she must maximize against the known probability α. What Proposition 1 demonstrates is that if α is sufficiently small relative to r, then the proposer behaves the same way she would in the standard ultimatum game—she opts to keep the entire good. When facing a Kantian respondent she will lose out, but she is willing to take that risk. However, if $\alpha \geq r$, then the proposer offers the smallest division that would not insult the dignity of the Kantian respondent. In this situation, the Hobbesian respondent benefits. He would have accepted an offer of less than r, but because of the threat posed by the Kantian respondent, he has received a larger offer. This latter fact illustrates the *interpersonal argument* discussed in the introduction: the Hobbesian respondent stands to benefit from the presence of Kantians. In so far as he is the respondent, the Hobbesian would not wish to convince others in his society of the correctness of the Hobbesian view because this would reduce α. Because the Hobbesian will not reject positive offers once they are made, the Hobbesian cannot emulate the Kantian in single-shot interactions. But the Hobbesian respondent can benefit from the presence of Kantians and the desire for greater benefit provides an incentive (a reason) to promote Kantian values in others.[7]

It might be objected that the force of this argument depends on the strength of the incentive that the Hobbesian has to promote Kantian values in others and that this cannot be fully evaluated without considering the fact that any profit the Hobbesian receives as a respondent from the presence of Kantians might be offset by the loss the Hobbesian incurs when he is the proposer.[8] Two responses to this objection are in order.

First, there is an important asymmetry between the two conceptions of self-worth that we are considering here. The Kantian never has even a prima facie reason to promote a Hobbesian conception of dignity in others. Rather, the Kantian wants all other persons to recognize their, and every other rational agent's, categorically superior value to 'things.' Although the Hobbesian does not recognize such a categorical difference in value, he stands to benefit in so far as he is the recipient in the game we describe, from the presence of those who hold such a view. The Hobbesian recipient thus has at least a prima facie reason to promote Kantian values in others in some circumstances.

Second, in a world of unequal power and wealth — precisely the circumstances the Hobbesian is keen to exploit for his own

[7] Hobbesians receive a benefit from the presence of Kantians in this game. Undoubtedly societies involve many different interactions, and whether or not Hobbesians would benefit in other strategic situations from the presence of Kantians is an open question which we cannot tackle here.

[8] We thank the Editor and an anonymous referee for pushing us on this point.

advantage — there will almost always be Hobbesians who have an all-things-considered reason to promote Kantian values. This is because whether the Hobbesian's prima facie reason translates into an all-things-considered reason depends on two variables: how often the Hobbesian expects to occupy the role of the proposer and the relative stakes that are on offer when the Hobbesian occupies each role. The severely disadvantaged Hobbesian is likely to find himself almost exclusively in the role of the recipient, especially when dealing with more advantaged parties where the stakes are more lucrative. Hobbesians in this situation have an all-things-considered reason to promote Kantian values in others. Given current dramatic inequalities, where roughly two billion people live in extreme poverty, Hobbesians among the least advantaged would have strong reasons to promote Kantian values in others.

The same holds for Hobbesians who are middle-class entrepreneurs and occupy the role of the proposer when dealing with parties of a lower economic status and the role of the respondent when interacting with wealthy members of the upper class. Even if such Hobbesians occupy the role of proposer seventy percent of the time, they would still stand to benefit from the presence of Kantians as long as their interactions with wealthy members of the upper class are, on average, three times more lucrative than their interactions with less affluent members of the lower class. In this situation, the benefits of Kantians in the population outweigh the losses that these Hobbesians face when they are in the position of the proposer.

Because the frequency with which the Hobbesian's prima facie reason to promote Kantian values translates into an all-things-considered reason depends on the size of the stake, and not simply on the frequency with which an agent is in the position of the respondent, our argument has a fairly broad applicability. Workers with varying degrees of skill or expertise are likely to be in the position of the respondent with high frequency. It may be that in a highly competitive global market, small and mid-sized firms are routinely in the position of respondent. Parties with stronger bargaining positions, such as highly skilled workers or larger firms, may alternate positions with greater frequency. But if the stakes of their interactions with affluent parties are sufficiently high, then the marginal losses that come from treating respondents as Kantians when they are proposers may be compensated for by the marginal gains they reap from their higher-stake negotiations with more powerful and affluent parties.

3.2 The intrapersonal argument

The previous argument assumes that there are basically two types of agents in the population of potential recipients, committed Hobbesians and committed Kantians. To construct our *intrapersonal argument*, we need to modify this assumption. As we note at the outset, the two conceptions of dignity that we describe here are at home within a variety of competing, and often conflicting, moral theories. One of the reasons that we may care about such moral theories is that we want to use them for decision making. That is, we hope that we might be able to utilize the substantive principles, weights, orderings, or other relationships laid out in such theories to improve the way that we handle difficult decisions. Moreover, the process of evaluating the various merits and shortcomings of such competing theories can take considerable time and there is nothing that guarantees that we are confronted with difficult choices only once we have established a firm intellectual and affective commitment to a single theoretical framework.

It is likely that the number of people who are fully committed to Hobbesian or Kantian evaluative frameworks will be smaller than the number of people who have been exposed to each but who have yet to fully commit to one. Our intrapersonal argument focuses on these agents. They have sufficient facility with each framework to reason from within them and they are capable of acting on reasons from each framework. Sometimes when such an agent acts as a Kantian, he reflectively endorses what he has done and affirms both his reasons for action and his conduct. Perhaps this is especially prominent when he interacts with friends and intimates. Sometimes such an agent acts for Hobbesian reasons and affirms the rightness of those reasons and actions. Perhaps this conduct is most frequent in competitive contexts such as market transactions, or when interacting with strangers.

On other occasions, such an agent may be personally torn as these standpoints conflict. The agent might act for Kantian (Hobbesian) reasons but experience feelings of regret that are grounded in a revaluation of his conduct from the Hobbesian (Kantian) standpoint. At different times, others, and even the agent himself, may view him as a weak-willed Kantian or a weak-willed Hobbesian, in that he wants to act on the basis of certain reasons but cannot bring himself to do so in practice.

To make our *intrapersonal argument*, we need only interpret α in the game above as representing the probability that the agent will act on Kantian, rather than Hobbesian reasons. This is still an instance of a single-shot game, as the parties do not have any expectation that they will encounter one another again in the future. Nevertheless, if the probability that the respondent's Kantian views hold sway is sufficiently high, the proposer will offer r to the respondent, who will accept.

When the agent reflects on this situation from the Kantian standpoint, his willingness to reject offers below r is endorsed as a recognition of the special status of his dignity. From this standpoint, the agent acts properly only when he acts for Kantian reasons. So the Kantian standpoint does not provide such agents with reasons to endorse either the conduct or reasons for action that emerge out of the Hobbesian perspective.

When the agent reflects on this situation from the Hobbesian standpoint, his willingness to reject offers below r on Kantian grounds is endorsed as a useful means of securing a larger payoff. As a result, the agent has Hobbesian grounds for recognizing that it would be less profitable to completely repudiate his Kantian commitments and convert fully to the Hobbesian camp.

This is an interesting result. When the conflicted agent encounters profitable but insulting offers, he experiences a kind of overlapping consensus of reasons. He finds the Kantian reasons for rejecting the offer compelling, because the Kantian's notion of dignity has some rational purchase on him. He can also see that being committed to this conception of dignity may also have a monetary advantage. This may not be sufficient to resolve the agent's conflict by converting him to Kantianism. But such an agent now has both Kantian and Hobbesian reasons *not* to convert completely to the Hobbesian perspective.

One might object that whether the conflicted agent has Hobbesian grounds to maintain his Kantian commitments will depend on the frequency with which those commitments materialize when the agent is the proposer, thus depriving him of offsetting benefits. However, α here only refers to the probability that the agent will act on Kantian reasons as a recipient. Our *intrapersonal argument* need not say anything about how the agent behaves as a proposer.[9] We think that it is quite reasonable to assume that conflicted agents might approach these situations somewhat differently.

[9] Even if we suppose that α represents how frequently the agent acts on Kantian reasons as either a proposer or a respondent, it would still be the case that the many conflicted agents will have overriding reasons not to reject their Kantian commitments because they are more frequently in the position of the respondent, or because the stakes are higher when they are respondents than when they are proposers.

One might also object that conflicted agents do not have Hobbesian reasons to endorse sometimes acting on Kantian reasons. Rather, all such respondents must do is convince the proposer that they are sometimes a Kantian, and this might be done by lying or by careful acting. Two responses to this objection are in order.

First, this objection effectively begs the question against the argument that we present. The reason is that in the static game α is not something that the agent chooses. It represents the disposition of the agent to choose in accordance with one set of considerations or another. The agent only has an imperfect disposition to choose as a Kantian because although the agent finds the Kantian view sufficiently compelling that he will act on it from time to time, he is not persuaded that the framework as a whole represents the best ethical theory. The intrapersonal argument is about the degree to which an agent, conflicted between these two sets of reasons, has grounds to endorse or repudiate the disposition for choice that would come from wholly adopting one of these competing standpoints. Although there do not seem to be reasons from the Kantian standpoint to maintain a Hobbesian view of the agent's own dignity, the agent can recognize the value, from the Hobbesian standpoint, of having and acting on a Kantian conception of the agent's own dignity.

Second, because the objection misrepresents our claims about what is going on in this case, it is an objection against a different game than the one presented here. It misrepresents our claims because it presumes that the conflicted agent has only instrumental reasons to support a Kantian conception of dignity. Only then does it become credible to think that those instrumental reasons support acting like a Kantian, rather than endorsing Kantian values.

This argument also distinguishes us from so-called 'two-tier' views like rule-egoism or rule-consequentialism (see Kagan 1998). A two-tiered approach to this problem might argue that the Hobbesian does best by adopting the rule 'refuse offers below a certain threshold.' If the Hobbesian could *credibly* commit to such a rule *in view of the proposer*, then he certainly would do better. The problem with such positions is that it is difficult to make such commitments credible in the single-shot game. If one adopts this rule for Hobbesian reasons, then when one is actually confronted with a low proposal one has Hobbesian reasons for discarding the rule and accepting the offer. The purpose of adopting the rule was to secure a higher offer. In the single-shot game, however, once an offer is announced, the Hobbesian has no reason to follow the rule. A smart proposer will know this, and thus will ignore any allegiances the Hobbesian declares to such a rule.

Convincing the proposer that one will stick to such a rule may not be a simple matter. Merely claiming to have adopted the rule is, in the parlance of economics, just cheap talk. The proposer should recognize that even those who will not stick to the rule have an incentive to claim that they will. Thus the proposer will ignore any claims that are not backed by something that makes them credible.

We are not claiming that Hobbesian reasoning will, if taken up a level, endorse Kantian conceptions of dignity. Instead, we are arguing that someone who is torn between the two foundational conceptions has Kantian reasons to reject profitable but insulting offers and has Hobbesian reasons not to resolve the disagreement over foundational theories in favour of the Hobbesian conception of dignity.

Our account also differs in some interesting ways from approaches to the ultimatum game that rely on precommitment. Many accounts have relied on the idea that an individual respondent might benefit from the threat of irrationality (see, for example, Frank 1988). If two agents are playing chicken — driving their cars at one another in order to see who 'chickens out' first by turning away — then it might be rational for one party to simply throw his steering wheel out of the car once it is moving, showing the other party that both of their lives are in the other party's hands. Throwing away the steering wheel is a kind of pre-commitment device. The agent takes this action at time t in order to prevent herself from being able to decide what to do at $t+1$. The rational agent chooses to deprive herself of the ability to rationally choose how to act at $t+1$ because doing so puts her in a more favourable strategic situation vis-à-vis the other player.

In our model, the Kantian does not act at a time t so as to prevent herself from being able to make a decision at time $t+1$. Rather, the Kantian would make the same decision at time $t+1$ as she does at time t. This is because the Kantian decides in accordance with her values at each point. Additionally, the precommitment argument (see Schelling 1966) usually focuses on the benefit of precommitment for the person who makes the commitment. Even if we were to interpret the Kantian's moral values as a kind of precommitment device, one of our results is that it is Hobbesians who benefit from the use of this device (that is, Hobbesians benefit from Kantians acting on Kantian reasons) and, therefore, that Hobbesians have an incentive to perpetuate its use (by cultivating the willingness of others to make Kantian choices for Kantian reasons).

One common story about precommitment distinguishes at least two faculties that influence the agency of persons, one that is slow and calculating (the rational faculty) and another that is quick and highly valenced (the emotions). The potential for an agent's emotional reactions to diverge from their rational valuations has given rise to significant discussion of the ways in which the emotions can benefit or harm decision-makers. For instance, one might argue that having a fiery temper, and having this disposition be publicly known, could serve as a kind of precommitment device in the sense that proposers would know that offers that evoke the recipient's anger might be rejected in the heat of the moment (Frank 1988).

Unlike this work, we are not committed to substantive claims about a psychological division of labour. It is perfectly consistent with our view that Kantian agents remain cool and calculating throughout the economic transaction. Nor are we committed to the idea that the agent's emotional reaction diverges from, or would conflict with, their calm and rational assessment of the situation.

Up to this point, the game that we describe is a single-shot interaction. Our first two arguments therefore ground reasons that support rejecting profitable but insulting offers that do not rely on the prospect that the agents will interact again in the future. As a result, our arguments differ significantly from results that look at ways in which iterated or repeated interactions can enable agents in a weak bargaining position to secure greater benefits.

In the next section we show how repeated interactions for our version of the ultimatum game lead the Hobbesian to imitate the Kantian by actually refusing offers below r.

4. The dynamic ultimatum game

We proceed now to study the game that is induced by the finite repeated iteration of the variation of the ultimatum game we introduced above. Now we suppose that a single proposer is facing the same respondent multiple times. In order to consider this situation, we must introduce some more formal assumptions about the proposer's reasoning. We will assume as before that the proposer is a Hobbesian who knows ahead of time the value of α and r, but does not know whether this particular respondent is Hobbesian or Kantian. To this we now add that the proposer knows (or is capable of inferring) how both the Hobbesian and Kantian respondents will behave and uses this information to infer which type she is facing. She reasons by use of Bayes's theorem to update her prior, α, on the basis of the actions she observes.

In game theory, the appropriate equilibrium concept for such situations is Sequential Equilibrium.[10] A Sequential Equilibrium comprises a set of strategies and a system of beliefs. Equilibrium strategies need to be *sequentially rational*, that is, optimal from the point of view of each information set at which a player is called upon to act, given the beliefs she maintains, and *consistent*, that is, compatible with Bayes's rule. The precise definition of consistency is rather tedious. Let us simply note that, in our context, as long as beliefs are derived using Bayes's rule they constitute consistent beliefs.

Let us begin with a brief note on the implications that Bayesian updating has for the nature of the equilibrium. It is possible for an agent to effectively reveal herself to the proposer as a Hobbesian. For example, if the proposer begins by offering a split which leaves the respondent with less than r, and the offer is accepted, the proposer now knows with certainty that the respondent is a Hobbesian, since a Kantian would refuse any offer below r no matter what the structure of the game. This might harm the Hobbesian in the long run, since now the responder can offer very little to the Hobbesian without fear of refusal.

Things are more complicated if an offer below r is refused by an unknown respondent. Let us imagine that the game is repeated twice. In period 1 the proposer plays some quantity $x > 1 - r$ and observes that his offer has been refused. When called to play again in period 2 he uses Bayes's rule in order to assess the probability that the recipient is a Kantian conditional on what has transpired so far. He is interested in computing $P(K|\{x, \text{'no'}\})$, the probability he is facing a Kantian respondent after his offer of $x > 1 - r$ was rejected in period 1. Let us

[10] The notion is due to Kreps and Wilson (1982a). Our vocabulary follows Osborne and Rubinstein 1994, Ch. 12. There is an extensive debate about the normative significance of game theoretic analysis in these sorts of situations (see, for example, Kadane and Larkey 1982, Kadane and Seidenfeld 1992). Given the assumptions we make about the knowledge structure of the situation, we believe that the game-theoretic equilibrium makes normative recommendations to the players *in this situation*. Whether these recommendations hold for other similar games or other situations of knowledge in this game, we leave for future research.

denote by β_i the probability with which a recipient of type $i \in \{K, H\}$ plays 'no' at stage 1. Using Bayes's rule we obtain

$$(1) \quad P(K|\{x, \text{'no'}\}) = \frac{1 \times \beta_K \times \alpha}{\beta_K \times \alpha + (1-\alpha) \times \beta_H}$$

and recognizing that a sequentially rational Kantian recipient always rejects an offer such that $x > 1 - r$ (therefore $\beta_K = 1$), (1) becomes

$$(2) \quad P(K|\{x, \text{'no'}\}) = \frac{\alpha}{\alpha + (1-\alpha)\beta_H}$$

Using (2) we observe that the degree of uncertainty depends on the Hobbesian recipient's strategy. If $\beta_H = 1$, that is, the Hobbesian will refuse any offer below r (he will behave like the Kantian), there is no updating over the priors. If $\beta_H = 0$ and the Hobbesian will accept any offer whatsoever the uncertainty is resolved.

So now we must decide what is best for a Hobbesian to do in the face of an offer below r in any but the last stage of the game. If he accepts the offer, he reveals his type and relegates himself to very small offers in the subsequent rounds. On the other hand, if he refuses, he is taking the worst of it on this round.

The analysis of this game is presented in Proposition 2. Our analysis draws from the work on reputation effects by Kreps and Wilson (1982b) and Milgrom and Roberts (1982).

Proposition 2: There exists a Sequential Equilibrium of the game repeated over two periods such that:

(1) if $\alpha \geq r$, then

R_H : In period 1, 'yes' for each $x \in [0, 1-r]$, 'no' otherwise
In period 2, 'yes' for each $x \in [0, 1]$

R_K : In both periods, 'yes' if $x \in [0, 1-r]$, 'no' otherwise

P : In both periods, $x = 1 - r$

(2) if $\alpha < r$, then

R_H : In period 1, 'no' with probability $\dfrac{\alpha(1-r)}{(1-\alpha)r}$, if $x \in \hat{I}(1-r, 1]$,
'yes' otherwise
In period 2, 'yes' for each $x \in [0, 1]$

$\mathrm{R_K}$: In both periods, 'yes' if $x \in [0, 1-r]$, 'no' otherwise

P : In period 1, $x = 1 - r$, if $\alpha \geq r^2, x = 1$ otherwise
In period 2, $x = 1 - r$ if he was rejected in period 1,
$x = 1$ otherwise

Proof:
Case 1: $\alpha \geq r$. We appeal to backward induction. The expected payoff of the proposer at stage 2 is

$$\mathcal{E}^2(x) = \begin{cases} x & \text{if } x \leq 1 - r \\ (1 - \frac{\alpha}{\alpha + (1-\alpha)\beta_H})x & \text{if } x > 1 - r \end{cases}$$

Consider a Hobbesian recipient who rejects with probability β_H any offer $1 - x$ that is less than r in period 1. With probability $1 - \beta_H$ he accepts $1 - x$, but then obtains zero in period 2, as he has revealed himself as a Hobbesian. With probability β_H he obtains zero in period 1, but retains the opportunity of gaining more than zero in period 2. Let us suppose for the moment that

$$(3) \quad \frac{\alpha}{\alpha + (1-\alpha)\beta_H} \geq r$$

Condition 3 implies that the Hobbesian recipient faces a *total expected payoff* equal to $(1 - \beta_H)(1 - x) + \beta_H r$. He maximizes by setting $\beta_H = 1$. Since $\alpha \geq r$ we know that condition 3 holds. The Hobbesian recipient refuses any offer that leaves him with less than r, thus causing the proposer to maintain her original beliefs in period 2. The proposer understands that in period 1 any proposal such that $1 - x < r$ will be refused for sure. Hence she proposes $1 - r$. In period 2, maximizing $\mathcal{E}^2(x)$ under the assumption that $\alpha \geq r$, she once more, she offers r to the recipient.

Case 2: $\alpha < r$. Setting $\beta_H = 1$ is not optimal in this case. If the Hobbesian recipient did so, the proposer, as above, would retain her original beliefs, although, unlike the previous case, this would lead the proposer to offer zero to the recipient in period 2. Thus, rejecting with certainty offers below r is not the Hobbesian recipient's best response. Decreasing β_H has the effect of increasing the probability the proposer assigns to the recipient being a Kantian after having observed a refusal. However, as long as β_H is such that

$$\frac{\alpha}{\alpha + (1-\alpha)\beta_H} < r$$

it cannot be part of the equilibrium strategies. It amounts to rejecting something positive, without making up for it in period 2. Alternatively, the Hobbesian recipient could choose β_H low enough so that

$$\frac{\alpha}{\alpha + (1-\alpha)\beta_H} > r$$

This is also sub-optimal. Increasing slightly β_H from that value would preserve condition 3 and, moreover, would increase the expected payoff $(1-\beta_H)(1-x') + \beta_H r$ Therefore, the Hobbesian recipient's best response entails

$$\frac{\alpha}{\alpha + (1-\alpha)\beta_H} = r \text{ or } \beta_H = \frac{\alpha(1-r)}{(1-\alpha)r}$$

Compared to case 1, faking the Kantian is costlier, and, therefore, it occurs less often.

Having determined the recipients' best responses as a function of the offer extended let us turn to the proposer. In period 1, the proposer maximizes

$$\mathcal{E}^2(x) = \begin{cases} x & \text{if } x \leq 1-r \\ (1-\beta_H)(1-\alpha)x & \text{if } x > 1-r \end{cases}$$

This implies that in period 1, $x = 1 - r$ if

$$(1 - \frac{\alpha(1-r)}{(1-\alpha)r})(1-\alpha) \leq 1 - r (\equiv ly\ \alpha \geq r^2)$$

or $x = 1$ otherwise. In period 2, he offers r if he observed a rejection in period 1 and 0 otherwise. ■

In the repeated game we have retained the *intrapersonal argument* and the *interpersonal argument*. Responders still fare better when α is large relative to r. In fact, the argument is made stronger. In the single-shot game if $\alpha < r$ the Hobbesian did equally poorly regardless of the value of α. In the one-shot case, the presence of a few Kantians was not sufficient to improve his situation. In the repeated game, there is a new possibility. If $\alpha < r$ but $\alpha > r^2$, the Hobbesian does better than he would if $\alpha < r^2$. (He would fare still better yet if $\alpha \geq r$.) So now he has some additional reason to prefer the presence of Kantians.

Beyond this, we have now developed the formal arguments necessary to underwrite the *emulation argument*. Both equilibria require the

Hobbesian recipient to refuse an offer that violates the *Kantian's* dignity in period 1, either with certainty (first case) or with some positive probability (second case). The Hobbesian recipient mimics the Kantian in an effort to obscure information, that is to say, in order to manipulate the proposer's assessment. Therefore, contrary to the static game, in the dynamic game the Hobbesian recipient acts on the opportunity that the presence of the Kantian presents and, by doing so, forces the proposer to make concessions.

Although it involves repetitions of a single game, our argument is significantly different from those presented to 'solve' the repeated Prisoner's Dilemma. Among some scholars, the latter game is almost synonymous with game theory and many political philosophers are attracted to the way that it illustrates a conflict between beneficial social outcomes that can only be achieved through cooperation and the direct pursuit of greater individual benefits. In the one-shot Prisoner's Dilemma, this conflict leads agents to choose in a way that produces worse social and individual outcomes. However, there is a substantial literature demonstrating how repetition of the Prisoner's Dilemma under certain circumstances enables agents to achieve better joint outcomes (Mailath and Samuelson 2006).

The game that we describe does not model a conflict between the social good and individual self-interest. It models two different conceptions of self-interest, neither of which is placed into conflict or tension with socially beneficial outcomes. Additionally, the story that unfolds here does not involve reciprocity (making it dissimilar to the repeated Prisoner's Dilemma). The proposer retains his negotiation advantage across both periods. Her ability to extend take-it-or-leave-it offers is not at stake. What may potentially vary, depending on the recipient's actions, is the intensity of the informational asymmetry. In some sense, along the equilibrium path the Hobbesian recipient commits to preserving the uncertainty, or, rather, acts so as to create a 'Kantian' reputation that in turn affects the proposer's decisions. In the case that $\alpha \geq r$ the threat never materializes. In period 1, the Hobbesian recipient is committed to reject any offer that leaves him with an amount smaller than r. At equilibrium, though, he is made an offer that he accepts. The proposer responds to the threat. He perceives it as a credible one, and, thus, makes an offer of precisely r, the smallest offer that will not get him in trouble.

Iterating the game with the same proposer makes it possible for Hobbesians to act as though they embrace a Kantian conception of dignity. Nevertheless, the Hobbesian still has reason to promote

Kantian values in others because he can act like a Kantian only if there are genuine Kantians in the population.

In the case of an agent who is conflicted between Hobbesian and Kantian conceptions of dignity, the iterated argument takes on greater force. In the *intrapersonal* argument, the conflicted agent had Kantian reasons to reject beneficial but insulting offers and Hobbesian reasons for maintaining at least a partial commitment to the Kantian values that make such behaviour possible. The conflicted agent retains these reasons when facing the prospect of repeated interactions with the same proposer. But in this new context, the conflicted agent also acquires Hobbesian reasons to reject insulting offers. That is, such an agent has both Kantian and Hobbesian reasons to reject the offer, both Kantian and Hobbesian reasons to maintain a commitment to the Kantian conception of dignity, and both Kantian and Hobbesian reasons to try to foster Kantian views in others.[11]

Even if the Kantian is not satisfied with the outcome of the argument we have made so far, and even if the Hobbesian has not been converted to Kantianism, it is worth pressing a bit more firmly on the degree to which the Hobbesian remains a Hobbesian. In particular, as a result of our interpersonal argument, our Hobbesian realizes that it is in his best interest not to attempt to convert anyone else to his original way of thinking. Because of the benefit he gets from the uncertainty introduced by the game, he further realizes that he should never admit to anyone that he is a Hobbesian himself. He should, when questioned, give Kantian responses to questions of evaluation and do the best he can to convince others of his belief in the Kantian's conception of dignity. When playing the repeated game, he will behave much like the Kantian.

On most operational levels, the Hobbesian would be indistinguishable from the Kantian. If there is a difference, it will be counterfactual. That is, the Hobbesian may affirm that under different circumstances, in which it is common knowledge that there are no Kantians for example, he might behave differently (although he would not even admit this to agents who might act as potential proposers for fear of revealing his Hobbesian nature). But it is interesting to think about what might happen to our Hobbesian if his general circumstances do not change. Over time, for example, he may come to expect offers that

[11] For agents who are conflicted between these two viewpoints, there are thus powerful forces aligned against resolving the conflict in favour of the Hobbesian conception of dignity. Over time, in the right environment, such an agent may simply not see sufficient value in retaining Hobbesian commitments.

reflect his Kantian view of himself. If such an expectation leads to a sympathy for the Kantian conception of dignity itself, then the Hobbesian may develop the kind of conflict in values to which our intrapersonal argument applies. Even if this is not the case, the Hobbesian may have powerful reasons to raise his children as Kantians, or at least as agents who have been exposed to Kantian values and who are therefore likely to be conflicted in the way we have described here. This line of argument is interesting, not for its normative implications, but as a basis for further work into how Kantian intuitions might have arisen in certain populations. We will leave this line of argument for another occasion, however.

5. Wertheimer and the social planning argument

So far we have examined a static and a dynamic version of the ultimatum game in order to illustrate the way that different conceptions of dignity can influence the behaviour of rational agents. We now turn to some implications of this work for policy. Before doing so, however, it is important to note that our project differs significantly from recent work that utilizes principles of fair bargaining to draw conclusions about norms of justice or fairness. In particular, much of that work takes place within the social contract tradition. Rawls (1971), Harsanyi (1975), and Gauthier (1986), for example, share the ambition of grounding a robust set of ethical and political norms in the fact that those norms would be agreed upon by rational agents who were placed in a position of trying to decide together the norms that ought to govern or regulate social interactions.[12]

In contrast, the interactions that we model do not involve bargaining over the rules or norms that ought to govern or constrain social behaviour. In the emulation argument, our agents are considering rules to guide their bargaining over monetary or material benefits rather than using bargaining to establish rules. As we will now

[12] They differ in the way they model the relevant choice situation, and in their assumptions about features of the contractors. For example, Rawlsian agents must decide on rules to govern the basic structure of society from behind a veil of ignorance that deprives them of information about their gender, social status, or other features that might bias them in favour of one social group. Harsanyi's agents do not decide behind a veil of ignorance, but they are modelled as having impartial preferences. In contrast, Gauthier seeks to ground moral principles as what would result from a bargain between real agents, were they to consider what the constraints on their interactions ought to be.

argue, the arguments we have made so far are relevant to social policy, but their relevance does not presuppose larger commitments to the social contract tradition.

Although the model that we present here is highly abstract, interactions that have this general structure occur routinely in actual practice. In some cases, these interactions are highly charged because they involve offers that entail some risk to the recipient, or breach a norm of propriety or cross a moral boundary. To be clear, we are not here considering cases where the offer involves imposing a net harm on the recipient, or involves the use of fraud, force, or deceit. Rather, we consider only cases in which both parties freely enter into a mutually beneficial transaction but in which there might exist grounds for questioning the moral acceptability of the transaction nonetheless. In particular, the charge that some such transactions are exploitative of the weaker party is commonly used to indicate that what may in all other respects be beneficial transactions suffer from a serious moral defect. Whether a particular interaction represents an instance of exploitation, however, depends, in part, on one's account of the moral wrong involved in such offers.

Kantian accounts of exploitation rely heavily on the claim that certain highly unequal transactions show a lack of moral respect for the disadvantaged party. For example, Ruth Sample has argued that 'exploitation involves interacting with another being for the sake of advantage in a way that degrades or fails to respect the inherent value in that being' (Sample 2003, p. 57). Sample articulates clearly the core intuition that drives many Kantian accounts of exploitation, namely that 'other human beings possess a value that makes a claim on us' and that when we exploit others we 'fail to honour this value in our effort to improve our own situation' (Sample 2003, p. 57; see also Siegel 2008, Wood 1997).

In contrast, Alan Wertheimer has articulated what might be viewed in the present context as a more Hobbesian account of exploitation. On Wertheimer's view, the central wrong committed in some voluntary but mutually beneficial transactions is not disrespect or degradation, but the fact that the more advantaged agent was able to leverage his or her bargaining situation to extract an unfair share of the surplus generated by the interaction (Wertheimer 1996, pp. 21–8). What the recipient of the deal is morally entitled to is determined, not directly by the needs or moral status of the agent, but by the share of the surplus that the recipient could have captured under more ideal market conditions. The test for exploitation, on this account, is whether the recipient would have been able to receive a larger share of the surplus in a more competitive or more ideal market.

Although these camps may differ in their account of which transactions are exploitative, they agree that exploitative interactions are morally wrong. As Wertheimer points out, however, the fact that a transaction is exploitative, and therefore morally wrong, does not necessarily entail that others should take affirmative action to prohibit such transactions. The reason is simply that, despite their moral taint, such transactions may represent the best alternative that is open to the recipient. If preventing the recipient from accepting such an offer does not improve her situation in any material respect, and in fact leaves her worse off, then intervention may even be counter-productive.

The fact that exploitation may be better, in a morally relevant sense, than neglect, poses a moral dilemma. To the extent that Kantians take actions to be the primary focus of moral evaluation, it may be easier for them to embrace a policy of prohibition on the grounds that sanctioning disrespectful or degrading conduct is worse than accommodating the liberty of agents to pursue their own life projects, even if this results in the unintended consequence that vulnerable populations suffer from a lethal neglect.

More moderate Kantians, like Sample, feel the force of this dilemma and are more likely to side with more permissive social policies of the form that Wertheimer advocates (see Sample 2003, pp. 86–7). We can capture the more permissive position concerning the moral merits of intervention in cases of mutually beneficial but exploitative offers by what we call Wertheimer's Principle (WP):[13]

(WP) If A makes an offer to B, from which B will benefit, and to which B willingly agrees, then permitting their interaction cannot be morally worse for B than prohibiting it, even if the transaction is unfair, exploitative or unjust to B

For Wertheimer, this principle establishes what ought to be the default view of exploitative offers. If the strong presupposition in their favour

[13] (WP) is entailed by two claims that Wertheimer makes. 'Given the non-ideal background conditions under which people find themselves, there should be a very strong presumption in favour of principles that would allow people to improve their situations if they give appropriately robust consent, if doing so has no negative effects on others, and this even if the transaction is unfair, unjust, or exploitative' (Wertheimer 2008, p. 84). And what Wertheimer calls the 'Non-worseness claim' which 'maintains that it cannot be morally worse for A to interact with B than not to interact with B if: (1) the overall interaction or package deal is better for B than non-interaction, (2) B consents to the interaction, and (3) such interaction has no negative effects on others' (Wertheimer 2011, p. 259).

cannot be rebutted then they ought to be permitted and social authorities should even be 'prepared to enforce [their] terms — if doing so is necessary to facilitate such transactions' (Wertheimer 2011, p. 214).

The arguments that we present above provide strong Hobbesian grounds for Hobbesians who are most likely to be the recipients of such offers (those who are most often in the position of the recipient of such offers, and who face less lucrative stakes when they are in the position of the proposer) to reject (WP). As it applies to our model, (WP) only has bearing on Hobbesian respondents. Kantian respondents would not accept an offer of less than *r* even if doing so was permitted by law. Therefore, (WP) would never be relevant — Kantians would not willingly agree to exploitative offers. Instead, the condition must be evaluated by considering a Hobbesian population — one where individuals would, if offered, freely accept an offer of less than *r*.

For illustration, let us consider a population composed entirely of Hobbesian recipients that faces exploitative offers from wealthy foreigners. A social planner from this community who endorses (WP) will adopt a laissez-faire policy about what kind of offers members of the relevant population can accept. Under this regime, recipients will routinely receive the smallest division of benefits possible.

Now consider a social planner[14] who, in violation of (WP), institutes a policy that prohibits deals that provide recipients with a share of benefits less than *r*. Moreover, suppose that, given limitations on resources, the planner knows that such a policy could only be sporadically enforced. In particular, the social planner adopts a policy of randomly monitoring a proportion of transactions, α. As we demonstrate in Proposition 1 above, as long as α is greater than *r*, all respondents in the population will receive a share of benefits of size *r*, making them better off than under the laissez-faire policy mandated by (WP).

As a result, if the population of Hobbesian recipients, as modelled above, could vote on the matter, they would vote unanimously against the laissez-faire policy and in favour of the one that we describe. This is because they face a classic problem of collective action. If actually presented with the exploitative offer, each recipient would prefer a positive material gain to nothing. But for this very reason, each prefers not to face the exploitative offer in the first place. Regulation is

[14] We use the term 'social planner' merely as a convenient proxy for whatever political process is necessary to enact legislation or pass regulations in a community.

justified in such a case because it enables the population of agents to achieve a goal that they each desire but could not bring about on their uncoordinated, individual initiative. This latter fact illustrates our final argument, which we call the *social planning argument*.

To his credit, Wertheimer recognizes that considerations of this kind, which he labels 'strategic considerations,' represent the most credible counterexample to (WP) (Wertheimer 2011, pp. 216–17). But he worries that the scope of this objection may be relatively limited since the structure of the argument requires that the policy intervention not be so severe as to dissuade the more advantaged party from interacting with the disadvantaged party on better terms. As such, he thinks the applicability of this objection hinges critically on the facts of the matter. And in some cases, such as research that is funded by pharmaceutical companies or other entities from high-income countries and carried out in comparatively disadvantaged populations of low- or middle-income countries, Wertheimer thinks that the prospects of prohibition working to the detriment of disadvantaged populations is particularly high. As such, he thinks the burden of proof lies with those who would reject (WP).

The results that we present here suggest, however, that what appear to the Hobbesian as strategic worries may be more endemic to such offers than Wertheimer's analysis recognizes. The problem is that, in its most general form, any exploitative offer represents a 'probe' that has the potential to reveal the agent's willingness to accept low offers. Following (WP) in one area may appear to eliminate inefficiencies that might be caused by the fact that pharmaceutical companies, for example, may not be interested in hosting clinical trials in some developing countries if they cannot do so on what amount to exploitative terms. By trying to capture the benefits of hosting such trials, policymakers in host countries reveal their willingness to accept exploitative offers. This, in turn, may alter their future ability to attract interactions that comply with norms of respect. Parties who already interact with such countries in other domains now have an incentive to be more aggressive in their bargaining in order to secure a larger share of the surplus of their interactions for themselves. Alternatively, some agents who are committed to making only respectful offers may simply be deterred from interacting with the host community. For example, many tourists may prefer to avoid destinations that have signalled a tolerance for exploitative activities, such as sex-tourism. Some may have Kantian motives, wishing to avoid complicity with regimes that sanction disrespect or degradation. Others may simply fear that such environments are commonly associated with a range of dangers that they would rather avoid.

Uniform policies prohibiting exploitative offers alleviate the problems associated with revealing one's willingness to be exploited. They do not of themselves ensure that the void of exploitative offers will now be filled with more beneficial and respectful alternatives. But in many cases, policy makers themselves can fill this void. For instance, minimum wage laws prohibit disrespectful wage schemes. Nonetheless they may also prevent some agents from finding employment. Rather than permitting exploitative employment practices, states can improve the status of the unemployed by enacting social welfare policies that enable the unemployed to meet their basic needs while exploring other employment, educational, or vocational opportunities.

In the case of international research, this may mean that regulations governing cross-national clinical trials should not be weakened to permit more exploitation. Instead, they might be supplemented by initiatives that reward private entities for conducting respectful research in low- and middle-income countries or that mandate state sponsors of research to promote such activities.

6. Conclusion

The strategy of this paper has not been to push the debate about competing conceptions of dignity back to moral foundations. It has been to model how agents with recognizably Kantian and Hobbesian conceptions of dignity respond to a choice situation in which an agent might perceive her dignity to be at stake and to consider the relative merits of those responses in light of those agent's own values. Our approach reveals some interesting asymmetries between the Hobbesian and Kantian evaluative standpoints in relation to their respective conceptions of dignity.

The Kantian standpoint has an important kind of stability. Kantian agents embrace their favoured conception of dignity as correct or true. They affirm this commitment publicly. In part, this is because they take this view to be normative for all rational agents. Kantians, that is, want others to recognize the status of every other rational agent as a member of the kingdom of ends and they want social policy to reflect this understanding.

The Hobbesian standpoint lacks this stability. The disconnect between the behaviour of Hobbesian agents and their considered conception of dignity goes beyond those inherent in most two-tiered normative theories.[15] In the static ultimatum game, the Hobbesian respondent would like to secure the more lucrative split of benefits that the Kantian achieves, but the Hobbesian respondent cannot achieve this on his own. In the static game, Hobbesian respondents benefit — not from acting like Kantians — but from the presence of real Kantians in the population. Only real Kantians will reject insulting offers and Hobbesian respondents benefit from the proposer's uncertainty about the recipient's type. So it is not that the Hobbesian has reasons to hide a Hobbesian justification for acting like a Kantian in the one-shot case. It is, rather, that Hobbesian respondents have reasons (of significant force for the most disadvantaged agents) for wanting others to be *actual* Kantians. If real Kantians did not exist, then some Hobbesians would have powerful reasons to invent them.

In repeated interactions, Hobbesians can emulate Kantian reactions to proposed divisions and thereby benefit from maintaining uncertainty about their type. But this result too depends on the presence of real Kantians in the population and on the Hobbesian publicly repudiating his considered view. Moreover, this repudiation would reach into the realm of policy, with Hobbesian agents preferring that social policy reflect a Kantian, rather than a Hobbesian, conception of dignity.

On a cognitive level, to fully accept or adopt a normative framework is to leave behind a state of uncertainty or conflict and to endorse the considerations that it generates as authoritative for deliberation and action. Kantianism supports this level of commitment, in that the agent who endorses this theory can view the move

[15] For example, Williams coined the term 'Government House' utilitarianism to refer to the way that utilitarians like Sidgwick held theories that required that the nature of the theory itself not be made public (Williams 1985). That is, Sidgwick claimed that utilitarian considerations could justify adopting and acting on rules or dispositions that diverge from what would result from a straightforward application of utilitarian reasoning to particular decisions. But in order for some agents to maintain a commitment to these rules or dispositions, they would have to remain ignorant of the true, utilitarian, justification. Although the sophisticated intellectuals of the 'government house' could grasp the way that non-consequentialist rules or dispositions could be grounded on utilitarian foundations, the 'commoners' would have to be shielded from such information.

Here, at least, the utilitarian is claiming that there is a path from a foundational utilitarian concern to a justification of seemingly non-utilitarian rules or principles but that it would be difficult to convey this explanation to regular folks in a way that will not cause bad consequences. The situation we describe is worse. Hobbesians must obscure their Hobbesian reasons for wanting others to be real Kantians and conflicted agents have both Hobbesian and Kantian reasons not to fully convert to the Hobbesian viewpoint.

from conflict to consistent endorsement as an improvement, from the standpoint of the theory itself. The agent has no reasons, grounded in the Kantian framework, for regretting this commitment.

This is not the case from the Hobbesian framework. Moving from a state of genuine conflict to fully accepting the Hobbesian theory would be regarded as *a mistake*, on Hobbesian grounds, by at least some Hobbesians. The Socrates of Plato's *Republic* treated the political community as the psyche of the agent writ large. In this case, the desire of some Hobbesians to perpetuate Kantianism among others in the community reflects the conflicted agent's Hobbesian reasons for preserving his genuinely Kantian conception of dignity.

The committed Hobbesian may treat these arguments simply as the consequences of Hobbesian commitments and the committed Kantian is likely to view them as insufficiently foundational. But for those of us who have yet to fully commit to one of these viewpoints, these considerations provide powerful reasons not to commit to the Hobbesian view and they provide some positive support for moving in the Kantian's direction.[16]

References

Barry, Brian 1989: *Theories of Justice*. Berkeley, CA: University of California Press.

Bicchieri, Cristina and Maria Luisa Dalla Chiara 1992: *Knowledge, Belief, and Strategic Interaction*. Cambridge: Cambridge University Press.

Binmore, Ken 1989: *Natural Justice*. New York, NY: Oxford University Press.

Copp, David 2006: *The Oxford Handbook of Ethical Theory*. New York, NY: Oxford University Press.

Cummiskey, David 1996: *Kantian Consequentialism*. New York: Oxford University Press.

Darwall, Stephen 2006: 'Morality and Practical Reason: A Kantian Approach'. In Copp 2006, pp. 282–320.

Frank, Robert H. 1988: *Passions within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton & Company.

[16] The authors thank Teddy Seidenfeld, Nicole Hassoun, and the participants of the Games and Decisions Working Group at Carnegie Mellon for their insightful criticisms. We also thank the Editor and three anonymous referees for valuable feedback.

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

Guth, Werner, Bernd Schwarze, and Rolf Schmittberger 1982: 'An Experimental Analysis of Ultimatum Bargaining'. *Journal of Economic Behavior and Organization*, 3, pp. 367–8.

Harsanyi, John C. 1976: *Essays on Ethics, Social Behaviour, and Scientific Explanation*. Dordrecht: D. Reidel.

Hawkins, Jennifer S. and J. Ezekiel J. Emanuel 2008: *Exploitation and Developing Countries: The Ethics of Clinical Research*. Princeton: Princeton University Press.

Hill, Thomas, Jr 1992: *Dignity and Practical Reason in Kant's Moral Theory*. Ithaca, NY: Cornell University Press.

——2003: *Respect, Pluralism, and Justice*. New York, NY: Oxford University Press.

Hume, David 1739: *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals,* second edition, ed. L. A. Selby-Bigge. Oxford: Clarendon Press, rprnt 1902.

Kadane, Joseph B. and Patrick D. Larkey 1982: 'Subjective Probability and the Theory of Games'. *Management Science*, 28, pp. 113–20.

Kadane, Joseph B. and Teddy Seidenfeld 1992: 'Equilibrium, Common Knowledge, and Optimal Statistical Decisions'. In Bicchieri, Chiara, and Luisa 1992, pp. 27–45.

Kagan, Shelly 1998: *Normative Ethics*. Boulder: Westview Press.

Kavka, Gregory S. 1986: *Hobbesian Moral and Political Theory*. Princeton, NJ: Princeton University Press.

Korsgaard, Christine M. 1996: *Creating the Kingdom of Ends*. New York, NY: Cambridge University Press.

Kreps, David 1982b: 'Reputation and Imperfect Information'. *Journal of Economic Theory*, 27, pp. 253–79.

Kreps, David and Robert Wilson 1982a: 'Sequential Equilibrium'. *Econometrica*, 50, pp. 863–94.

Levi, Issac 1986: *Hard Choices: Decision Making under Unresolved Conflict*. Cambridge: Cambridge University Press.

Mailath, George J. and Larry Samuelson 2006: *Repeated Games and Reputations: Long-Run Relationships*. Oxford: Oxford University Press.

Milgrom, Paul and John Roberts 1982: 'Predation, Reputation and Entry Deterrence'. *Journal of Economic Theory*, 27, pp. 280–312.

Murnighan, J. Keith 2008: 'Fairness in Ultimatum Bargaining'. in Plott and Smith 2008, pp. 436–53.

Neilsen, Kai and Robert Ware 1997: *Exploitation*. New York: Humanities Press.

Osborne, Martin J. and Ariel Rubinstein 1994: *A Course in Game Theory*. Cambridge, MA: MIT Press.

Plott, Charles R. and Vernon L. Smith 2008: *Handbook of Experimental Economics Results*. Amsterdam: North Holland.

Rawls, John 1971: *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Sample, Ruth J. 2003: *Exploitation: What it is and Why it is Wrong*. Lanham: Rowman and Littlefield.

Schelling, Thomas C. 1966: *Arms and Influence*. New Haven, CT: Yale University Press.

Sen, Amartya 2002: *Rationality and Freedom*. Cambridge, MA: Harvard University Press.

Siegel, Andrew W. 2008: 'Kantian Ethics, Exploitation, and Multinational Clinical Trials'. In Hawkins and Emanuel 2008, 175–205.

Thaler, Richard H. 1988: 'Anomalies: The Ultimatum Game'. *The Journal of Economic Perspectives*, 2, pp. 195–206.

Wertheimer, Alan 1996: *Exploitation*. Princeton: Princeton University Press.

—— 2008: 'Exploitation in Clinical Research'. In Hawkins and Emanuel 2008, 63–104.

—— 2011: *Rethinking the Ethics of Clinical Research*. Oxford: Oxford University Press.

Williams, Bernard 1985: *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Wood, Allen 1997: 'Exploitation'. Neilsen and Ware 1997, pp. 2–26.