Why the Angels Cannot Choose[1]

J. McKenzie Alexander

Abstract

Decision theory faces a number of problematic gambles which challenge it to say what value an ideal rational agent should assign to the gamble, and why. Yet little attention has been devoted to the question of what an ideal rational agent *is*, and in what sense decision theory may be said to apply to one. I show that, given one arguably natural set of constraints on the preferences of an idealised rational agent, such an agent is forced to be indifferent among entire families of goods, and hence cannot choose among them. This result illustrates the dangers of speaking of the choices of an 'ideal rational agent' when one does not make precise the exact nature of the idealising assumptions. The result may also be viewed as providing an upper bound on the kinds of idealising assumptions which can be made for rational agents, beyond which the very concept of choice becomes attenuated.

Keywords: Decision theory, rationality, idealisation

1. Introduction.

In 'Vexing Expectations', Nover and Hájek argue that the Pasadena game,[2] a variant on the well-known St. Petersburg game, poses a serious problem for decision theory. The problem being *how to value the game*. Unlike ordinary problems of decision theory, the Pasadena game was defined so that the series one would normally use as the expected value of the game is *conditionally convergent*, meaning that the order in which the terms are summed can determine the value of the series. Yet since there is no natural order in which to sum the terms,[3] the Pasadena game thus appears to be capable of having any value from negative infinity to positive infinity.

In what follows, I will not broach the question of what value, if any, should be assigned to the Pasadena game. I avoid this question for two reasons. The first reason is that, in some sense, the question has been settled by Fine [2008] who showed that it is consistent with the standard axioms of decision theory to assign any value whatsoever to the Pasadena game. The second

[1] Special thanks are owed to Franz Dietrich for suggesting the particularly elegant axiomatization underlying the proofs of section 3. In addition, I am grateful to Jeff Barrett, Richard Bradley, Erik Curiel, Alan Hájek and Brian Skyrms for their helpful comments on earlier drafts of this paper.

[2] The game is defined as follows: I flip a fair coin repeatedly until it lands heads. If it lands heads on the nth toss, the amount paid equals $\$(-1)^{n-1} \cdot \frac{2^n}{n}$. If the amount is positive, I pay you that amount of money, and if the amount is negative, you pay me.

[3] The decision problem specifies the payoff p_n received if the coin lands heads for the first time on the nth toss, but there is no reason why one must calculate the expected value of the game by computing $\sum_{n=1}^{\infty} \frac{1}{2^n} \cdot p_n$. Although if one does use that as the 'expected value' of the game, it turns out that the game has a value of $\ln(2)$.

reason is that, if one considers nonstandard (or perhaps extended) types of decision theory, Easwaran [2008] has shown that consideration of repeated plays of the Pasadena game enables one to assign the unique value of $\ln(2)$ to the game. Although the problem of valuing the Pasadena game is not, strictly speaking, solved, it does seem that Fine and Easwaran have taken us several steps towards a solution.

But consider the question of why we should take the problem of valuing the Pasadena game seriously in the first place. Why might we not be inclined to do so? Largely because ordinary decision theory only considers problems with finite state spaces and finite payoffs and the Pasadena game, in contrast, involves both an infinite state space and unbounded payoffs. Neither one of these assumptions makes sense in the real world, so the most natural solution to the problem is to reject the game. However, Nover and Hájek argue that none of the reasons given for getting rid of the Pasadena game are satisfactory.

Consider, for one, the claim some have made that decision theoretic problems *must* have finite state spaces.[4] What's so special about finite state spaces? Although it may be true, as a matter of contingent fact, that all real world decision problems have finite state spaces, Nover and Hájek argue that decision theory has, or should have, a loftier goal in sight. Decision theory, on their view, pertains to *ideal rational agents* as well as finite humans:

> Decision theory is not merely about humans, but is also about — perhaps even especially about — ideal rational agents, and it does not follow that our idealized theory of ideal rationality should be so constrained … we happily idealize away our all-too-human nature elsewhere in our theorizing — in economics, for example. And again, even decision theory and probability theory themselves ignore our finiteness in other respects, assuming as they do that we are logically omniscient, that our preference orderings are infinitely rich, and so on.
>
> [Nover and Hájek 2004: 247]

Consider also the claim that decision theoretic problems *must* have finite payoffs (or, alternatively, bounded utility functions). Again Nover and Hájek argue that we do not have good *conceptual* reasons for requiring this, even if it is true as a matter of contingent fact (italics mine):

> But even supposing that all human utility functions happen to be bounded, so what? Again, it does not follow that our idealized theory of ideal rationality should be so constrained. After all, we do not eschew [continuous] utility functions, even though as a matter of contingent fact, all humans have finite thresholds for their perception of reward. Much as we impose no bounds on the sensitivity of our discriminations of utility, we should impose no bounds on the utilities themselves. If we are prepared to idealize 'in the small', we should equally be prepared to idealize 'in the large', fully aware that in doing so we pay no heed to our contingent limitations, in both directions. *And even if we do acknowledge the limitations of human beings, decision theory is not specifically about human beings. It is about rational decision makers in the abstract.*
>
> [Nover and Hájek 2004: 248]

In both passages, Nover and Hájek appeal to an 'ideal rational agent' or a 'rational decision maker in the abstract'. Just what *is* an ideal rational agent? What can one do? How do we know it even makes sense to talk of such an entity as engaged in anything like *choice* in the

[4] This is considered by Nover and Hájek as one possible response to the problem of valuing the Pasadena game: restrict decision theory to finite state spaces. They reject it (and rightly so, I believe) as being untenable. Many decision theories, including that of Savage, allow for infinite state spaces.

first place? The question is worth considering, for if speaking of ideal rational agents is legitimate, it opens up a whole new realm of study for decision theory. Just as Cantor tamed the paradoxes of infinity by showing us how to treat them within set theory, perhaps the paradoxes of the Pasadena and St. Petersburg game may be tamed by a decision theory for ideal rational agents residing in Plato's heaven.

Speculating about the capabilities of ideal agents has an esteemed history in philosophy, but a history which also reveals that such speculation may lead to unexpected conclusions. Aquinas, in *Summa Theologica*, engages in this task in his 'Treatise on the Angels'. There we discover that angels not only may be located in a place (*QLII*, A1),[5] but that they may act (*QLVIII*, A1), and that they have free will and can choose (*QLIX*, A3). Yet choice among angels turns out to be very different from choice among persons. For one, angels lack what Aquinas calls the 'sensitive appetite' required to sin by choosing evil.[6] The appetitive faculty of the angels stems from the intellect, and governs their will differently (italics mine):

> [T]he act of the appetitive faculty comes of this, that the affection is directed to something outside. Yet the perfection of a thing does not come from everything to which it is inclined, but only from something which is higher than it. Therefore it does not argue imperfection in an angel if his will be not determined with regard to things beneath him; but it would argue imperfection in him, where he to be indeterminate to what is above him.
>
> (*QLXIII*, A1)

So the perfection of the angels is compatible with their will not being determined by any Earthly trifles beneath them, unlike us humans whose will is often determined by material goods (e.g., the pursuit of food, clothing, shelter). On the other hand, the angels, being perfect, have their will determined by that which is above them: God.

Aquinas knew a lot about angels. Nover and Hájek know a lot about decision theory. Through a remarkable accident of history there is a line of inquiry pursued by Aquinas which we must press upon them: when we speculate about 'rational decision makers in the abstract' will the decisions of these ideal rational agents look anything like ours? How, and what, would these idea rational agents choose? It is a question worth asking because, when people talk about ideal rational agents, quite often these ideal agents seem a lot like us, except perhaps having an unbounded utility function or being able to play a game infinitely many times.

Somewhat sadly, it turns out that the very concept of an ideal rational agent is altogether much more problematic than it first appears. In what follows, I will argue that according to one natural conception of an 'ideal rational agent', the applicability of the concept of *choice* becomes severely attenuated. Hence, as Davidson [1974] said, '[t]he trouble is, as so often in philosophy, it is hard to improve intelligibility while retaining the excitement.' More precisely I will show that if an ideal rational agent has an infinitely rich preference ordering which obeys relatively minor coherence criteria, then the agent must be indifferent between vast numbers of goods. Ideal rational agents cannot really be said to be capable of choosing, then, in any sense we would recognise — much like Aquinas' angels.

2. The ideal rational agent.

2.1 Constraints upon belief

[5] Here I adopt the convention of referring to portions of *Summa Theologica* by the question and article number. For example, (*QLII*, A1) indicates the discussion of question 52, 'Of the Angels in Relation to Place', and the first article, 'Whether an angel is in a place?'

[6] Note that this does not mean that angels cannot sin. They simply cannot sin by choosing an evil outcome (Aquinas uses the example of adultery). The fallen angels sinned by exercising their will to an inappropriate degree.

As previously noted, decision theory traditionally analyzes finite choice problems faced by finite rational agents. Thinking about choice problems faced by ideal rational agents first requires that we be precise about what this kind of agent *is*. An ideal rational agent is supposed to be just like an ordinary rational agent except that certain 'naturally occurring' constraints are relaxed. What constraints are these? Some which immediately suggest themselves as candidates for consideration are the following:

1. The number of beliefs that the agent can hold at one time.
2. The computational complexity of problems that the agent can solve (and the speed with which the agent may solve them).
3. The amount of utility that the agent can receive from any particular outcome.

Since finite rational agents like us appear to hold an infinite number of beliefs — even an uncountably infinite number of beliefs — it might not be clear what, exactly, relaxing the first constraint amounts to. But notice that there is something peculiar about the way in which we are capable of holding infinitely many beliefs: we are only able to do so in a virtual sense [Pettit 1995] through the use of linguistic tricks such as quantification or recursion.

Constraints (2) and (3) are more easily seen to be ways in which an ideal rational agent would properly differ from rational agents like ourselves. We all recognize that our computational powers are limited and slow and error-prone. Even here, though, we could conceive of a hierarchy of ideal rational agents of differing abilities. How? It turns out that the kinds of problems which can be solved computationally divide into a hierarchy of complexity classes [Arora and Barak 2009]. An ideal rational agent capable of instantaneously solving a computational problem of a given complexity class is less powerful than an ideal rational agent capable of instantaneously solving a problem belonging to a higher complexity class.

For the purpose of this paper, let us bracket this question by assuming that an ideal rational agent can instantaneously solve, at zero cost, any problem which has a solution. If it should turn out that the solution to a particular problem happens to depend on, say, which axiomatization of set theory one uses, let us assume that an ideal rational agent believes some extension of ZFC which settles that question.[7]

Finally, relaxing constraint (3) means that an ideal rational agent can have an unbounded utility function. This seems sensible, for the most compelling argument for requiring bounded utility functions when modelling actual rational agents derives from the fact that we have only finitely many mental and physical states. Hence any given choice problem really only involves finitely many options and, given the standard assumptions underlying von Neumann-Morgenstern utility theory, that can be captured by a bounded utility function.

Thus we have the following partial characterization of an ideal rational agent: (1) an agent who may have infinitely many actual beliefs, although need not; (2) an agent capable, in principle, of solving any computational problem instantaneously, but — since she may not know every truth — is not necessarily omniscient; and (3) an agent whose preferences are capable of being described by an unbounded utility function.

This partial characterization of an infinitary rational agent is incomplete since we have only considered what *epistemic* constraints may or may not apply to her beliefs. We now need to reflect upon what constraints apply regarding her preferences.

2.2 Constraints upon preference

[7] Can two ideal rational agents believe incompatible extensions of ZFC? Let us assume that they can. Although ideal rational agents are, according to Nover and Hájek, logically omniscient, that just means that they draw inferences from their set of beliefs correctly and without fail. It does not mean that they have only true beliefs. As the story of the Fall illustrates, even angels can have differences of opinion.

Let us begin by defining goods as arbitrary spatially extended objects with associated properties.

Definition 1. Let P be a set of properties. A *good* is subset of $\mathbb{R}^3$ where each point is assigned a subset of properties from P. More precisely, let $X \subseteq \mathbb{R}^3$ and let $f: X \to 2^P$, where '2^P' denotes the power set of P. Then the ordered pair $\langle X, f \rangle$ is a good. A set of properties assigned to a point of X is called a *bundle of properties.*[8]

Although bundles of properties are essential to the definition of a good, quite often it is unnecessary to mention them explicitly. In what follows, I shall sometimes refer to $X \subseteq \mathbb{R}^3$ as a good; in this case, it should be assumed that some assignment of properties to points exists, even though it is not mentioned.

It must be noted that this conception of goods only handles *intrinsic* properties of objects. *Relational* properties of objects give rise to special problems for they cannot easily be represented in the framework here. That said, in section 4.1 I offer some suggestions on how the conception of goods developed here might handle some kinds of relational properties. Yet I would like to note that nothing truly important hinges upon this claim: even if what I say only holds for choice problems concerning the intrinsic properties of goods, it still shows that decision theory for ideal agents is significantly different from decision theory for ordinary agents.

A rational agent's preferences should conform to some basic coherency criteria. Perhaps the most basic coherency criterion one could specify [see Elster 1985] concerns the satisfiability of the agent's set of desires: that is, a rational agent must have a consistent set of desires in that some possible world exists where all of the agent's desires are satisfied. Standard von Neumann-Morgenstern utility theory goes beyond this by requiring an agent's preferences to satisfy a variety of relations pertaining to lotteries (simple or compound) over basic goods. Let us consider two coherency requirements more akin to satisfiability in their simplicity than the complex probabilistic requirements underlying von Neumann-Morgenstern utility theory.

Preference invariance under isometry. Suppose I ask you to choose between a chocolate bar and an orange, both of which are resting on a table. Suppose further that you, in fact, prefer the chocolate bar. Now suppose I slide the chocolate bar two centimetres to the right and ask you to state your preferences again. It seems obvious that there is no reason why a simple translation of the good two centimetres to the right should affect your preference in any way whatsoever. Likewise, if you prefer the chocolate bar to the orange, you should continue to prefer the former to the latter even if I rotate the orange by 45 degrees.[9] Neither one of the rigid translations should affect anything relevant to the choice under consideration. Let's formalise this as follows. Let $\mathfrak{I}$ denote the set of isometries of $\mathbb{R}^3$ which preserve the standard Euclidean metric. If $G \subseteq \mathbb{R}^3$ is a good, the outcome obtained by transforming the good by some isometry $\varphi \in \mathfrak{I}$ will be denoted G^φ. Finally, let $\lesssim$ denote the preference relation over goods. (In what follows, the indifference relation will actually be used more often and is denoted by $\sim$.) The first proposed invariance requirement is the following:

Axiom 1. *For all goods* G *and all isometries* $\varphi \in \mathfrak{I}$*, it is the case that* $G \sim G^\varphi$.

Note that for purposes of notational simplicity I have suppressed any mention of the property bundles here. The idea is straightforward: any rigid transformation does not change the internal structure of the good, in the sense of which points are assigned to which property bundles: it only changes the location of the good in space.

[8] This terminology is useful because, later on, we will need to speak of every bundle of properties appearing in a good.

[9] Assuming, of course, that you know everything there is to know about the orange already. If the orange had a rotten patch which you only saw for the first time after the rotation, then it would be rational for your preferences to change as a result of the transformation.

A number of objections to this suggested axiom undoubtedly come to mind. For example, why should our preferences remain constant for a good that is transformed in any spatial direction? Presumably one would prefer that the compost pile in the garden remain where it is, rather than being shifted into the sitting room.

Although there is some intuitive force behind this concern, it is worth nothing that the intution does not hold true for all goods at all times. In particular, the assumption that people's preferences are basically constant under local transformations (or at least a class of local transformations) is needed by Thaler and Sunstein [2009] when they argue that rearranging the items in their 'restaurant' example respects and preserves the autonomy of persons. My point being that the requirement of preference invariance under isometry is understandably *controversial* but it is by no means completely unmotivated.

That said, more needs to be said in defence of axiom 1. These brief remarks are only intended to establish that, at least for some class of goods, preference invariance under isometry is not obviously false. In section 4, I shall provide additional arguments to motivate the claim that an ideal rational agent — at least under one idealization — would have preferences satisfying this axiom.

Preference invariance under mereological composition. Consider again the choice problem of a chocolate bar and an orange, both resting on a table. Suppose that I then ask you to choose between a chocolate bar and a loaf of bread, or an orange and a second loaf of bread identical to the first. Intuitively, the introduction of the loaf of bread should not change things: if you prefer the chocolate bar to the orange, then you should prefer the chocolate bar and the loaf of bread to the orange and the loaf of bread. This second proposed invariance requirement can be stated more formally as follows:

Axiom 2. *For all goods G_1, G_2, and G, if none of the goods overlap and $G_1 \sim G_2$, then $(G_1 \oplus G) \sim (G_2 \oplus G)$.*

The symbol '$\oplus$' denotes the operation of mereological summation.

As with preference invariance under isometry, this axiom is not uncontroversial. For example, suppose you are stranded on a tropical island and have to hunt animals for food. Given a choice between a baseball bat or an oar, you might well be indifferent between having a bat or an oar as both are equally good for clubbing slow-moving animals. Yet given the choice between the composite goods (bat $\oplus$ rowboat) or (oar $\oplus$ rowboat), you may well strictly prefer the latter combination over the former. If so, this constitutes a violation of axiom 2.

The example is compelling, yet the reason it compels depends upon its invoking an implicit assumption not part of the present framework. Axiom 2 serves as a kind of independence axiom, in the sense that it says adding the same object to two bundles of goods should not change the indifference relation concerning the original bundles. The reason being that, since the choice concerning G_1 or G_2 just depends upon the properties of G_1 and G_2, including G no matter what should not affect one's preferences because the properties of G_1 and G_2 are not changed. (Recall my earlier remark that the concept of a good developed here does not treat relational properties.) When the objects in each bundle interact so as to acquire new functions or causal capacities, as in the thought experiment, it is obvious that one's preferences might change. But the reason they change is because the properties of the goods, when combined, are not the same as the properties of the goods in isolation. Hence the alleged counterexample invokes a type of good which the present framework does not attempt to treat, and so should not be seen as undermining the plausibility of the axiom.

3. Why the Angels Cannot Choose.

First, some definitions.

Definition 2. Two goods $\langle X, f\rangle$ and $\langle Y, g\rangle$ are said to be *weakly pointwise property equivalent* if for every point $p \in X$ there exists a point $p' \in Y$ such that the bundle of properties assigned to p is identical to the bundle of properties assigned to p', and vice versa.

The idea of weak pointwise property equivalence is that the two goods feature the same bundles of properties, although perhaps in different degrees. It warrants the name *weak* pointwise property equivalence because it only requires that, given a bundle of properties found in one good, one is able to find a *single* point assigned the same bundle of properties in the other good.[10] For example, two alloys of gold and lead which contain different amounts of gold and lead are weakly pointwise property equivalent, because the only property bundles featuring in either alloy are the properties of being gold or being lead. However, a pure gold ingot and a pure lead ingot are not weakly pointwise property equivalent because one ingot only includes the property of being gold, whereas the other ingot only includes the property of being lead.

Definition 3. A good G is *pure* if only a single bundle of properties appears in G.

Definition 4. Let $G = \langle X, f\rangle$ be a good. We say that G is *bounded* if X is a bounded subset of $\mathbb{R}^3$.

Definition 5. A good G is *thick* if every bundle of properties appearing in G is assigned to a region of X having a nonempty interior.

Given the above, one can prove the following:

Lemma 1. *Let A be an ideal agent who has a consistent and coherent preference ranking $\lesssim$ defined over all possible goods satisfying axioms 1 and 2. If G_1 and G_2 are goods which are pure, bounded, and thick involving the same bundle of properties, then A must be indifferent between G_1 and G_2.*

Proof. Let $G_1 = \langle X, f\rangle$ and $G_2 = \langle Y, g\rangle$ be two goods which satisfy the assumptions. Because G_1 and G_2 are thick, X and Y are bounded subsets of $\mathbb{R}^3$ having nonempty interior. By the Banach-Tarski paradox (see Wagon, 1985), the sets X and Y are equidecomposible. That is, there exists a partition $X_1, \ldots, X_n$ of X, and a set of corresponding isometries $\varphi_1, \ldots, \varphi_n$ such that

$$X = X_1 \oplus X_2 \oplus \cdots \oplus X_n$$

and

$$Y = X_1^{\varphi_1} \oplus X_2^{\varphi_2} \oplus \cdots \oplus X_n^{\varphi_n}.$$

From Axiom 1, it follows that $X_1 \sim X_1^{\varphi_1}$. Axiom 2 then implies

$$(X_1 \oplus X_2 \oplus \cdots \oplus X_n) \sim \left(X_1^{\varphi_1} \oplus X_2 \oplus \cdots \oplus X_n\right).$$

Note that I am here invoking the convention described previously of referring to a good by simply referring to the set of spatial points it occupies. It should be clear from the construction how to obtain the function assigning bundles of properties to the points of $X_1^{\varphi_1} \oplus X_2 \oplus \cdots \oplus$

[10] A natural definition of *strong* pointwise property equivalence between two goods $\langle X, f\rangle$ and $\langle Y, g\rangle$ would require there to be a bijection $h: X \to Y$ such that, for every $x \in X$, the bundle of properties assigned to x was the same as the bundle of properties assigned to $h(x)$.

X_n. Namely, the points of $X_2 \oplus \ldots \oplus X_n$ are simply assigned the same property bundles that they are assigned in G_1. The points of $X_1^{\varphi_1}$ are assigned the naturally corresponding property bundles that X_1 is assigned in G_1, tranformed by the isometry φ_1.

Invoking Axiom 1 again, it follows that $X_2 \sim X_2^{\varphi_2}$. Applying Axiom 2 again, we see that

$$\left(X_1^{\varphi_1} \oplus X_2 \oplus \cdots \oplus X_n\right) \sim \left(X_1^{\varphi_1} \oplus X_2^{\varphi_2} \oplus \cdots \oplus X_n\right).$$

Transitivity of the indifference relation yields

$$(X_1 \oplus X_2 \oplus \cdots \oplus X_n) \sim \left(X_1^{\varphi_1} \oplus X_2^{\varphi_2} \oplus \cdots \oplus X_n\right).$$

Continuing in this fashion n times we obtain

$$(X_1 \oplus X_2 \oplus \cdots \oplus X_n) \sim \left(X_1^{\varphi_1} \oplus X_2^{\varphi_2} \oplus \cdots \oplus X_n^{\varphi_n}\right)$$

that is, $G_1 \sim G_2$. ■

The real magic of the above proof lies in its use of the Banach-Tarski Paradox. The 'paradox' receives its name largely due to the counterintuitive nature of the claim, but it is a proper theorem of Zermelo-Frankel set theory with the axiom of choice. The simplest form of the theorem states that any two solid spheres in $\mathbb{R}^3$ are equidecomposible: one sphere may be divided into a finite number of pieces[11] which can be rearranged using only rigid motions — translations and rotations — in order to form a solid sphere of any size whatsoever.[12]

One should note that the above proof does not actually assume that ideal rational agents are capable of performing Banach-Tarski style manipulations. Rather, the requirements imposed by axioms 1 and 2 on the indifference relation of an ideal rational agent force the ideal agent to be indifferent between G_1 and G_2 given the mere *possibility* of Banach-Tarski style manipulations being performed.

Although lemma 1 only concerns pure, thick and bounded goods, we can generalize this result. Before we do this, though, we need to first establish that the indifference relation is preserved under mereological sums.

Lemma 2 (Preservation of Indifference). *Let G_1, G_2, G_3 and G_4 be nonoverlapping goods. If $G_1 \sim G_3$ and $G_2 \sim G_4$, then $(G_1 \oplus G_2) \sim (G_3 \oplus G_4)$.*

Proof. Since $G_1 \sim G_3$, it follows from Axiom 2 that $(G_1 \oplus G_2) \sim (G_3 \oplus G_2)$. Similarly, since $G_2 \sim G_4$, it again follows from Axiom 2 that $(G_2 \oplus G_3) \sim (G_4 \oplus G_3)$. Transitivity of the indifference relation yields $(G_1 \oplus G_2) \sim (G_3 \oplus G_4)$ since the operation of mereological summation is commutative. ■

At this point we can prove the following generalization of lemma 1.

Theorem 1. *Let A be an ideal agent who has a consistent and coherent preference ranking $\lesssim$ defined over all possible goods satisfying axioms 1 and 2. If G_1 and G_2 are thick and bounded goods which are weakly pointwise property equivalent and feature only finitely many property bundles, then A must be indifferent between G_1 and G_2.*

[11] The minimum number is five [see Wagon 1985: 40].

[12] This theorem is remarkable, in part, because rigid motions are *measure preserving*. How is it possible for a ball to be decomposed into five pieces which, through measure preserving operations, becomes a ball twice the original size? The secret lies in the fact that the five pieces into which the ball is decomposed are non-measurable sets.

Proof. Let $G_1 = \langle X, f \rangle$ and $G_2 = \langle Y, g \rangle$ be thick goods which are weakly pointwise property equivalent, and let $p_1, \dots, p_n$ enumerate the property bundles appearing in both goods. Let $G_1(p_i)$ denote the subgood of G_1 which only involves the bundle of properties p_i. That is,

$$G_1(p_i) = \{x \in X : f(x) = p_i\}.$$

Similarly, $G_2(p_i)$ denotes that subgood of G_2 which only involves the bundle of properties p_i.

From lemma 1, we know $G_1(p_i) \sim G_2(p_i)$ for all i. Repeated application of Lemma 2 yields

$$\left(\bigoplus_{i=1}^{n} G_1(p_i) \right) \sim \left(\bigoplus_{i=1}^{n} G_2(p_i) \right)$$

which is the same thing as $G_1 \sim G_2$. ■

Examining the proof of theorem 1 reveals why the assumption that there be only a finite number of property bundles is needed: repeated application of lemma 2 only ensures that the indifference relation is preserved under a finite number of mereological sums. In some sense, this seems to be a slightly artifical restriction. Since we are considering the preferences of ideal rational agents who are 'logically omniscient' with 'infinitely rich' preference orderings, why should the indifference relation not be preserved under countably many mereological summations, or even uncountably many? Consider, then, the following axiom:

Axiom 3 (Countable preservation of indifference). *Let $G_1, G_2, \dots$ and $H_1, H_2, \dots$ be countable families of nonoverlapping goods. If $G_i \sim H_i$, for all i, then*

$$\left(\bigoplus_{i=1}^{\infty} G_i \right) \sim \left(\bigoplus_{i=1}^{\infty} H_i \right).$$

Axiom 3, combined with the previous results, enables us to prove a further generalisation.

Theorem 2. *Let A be an ideal agent who has a consistent and coherent preference ranking $\lesssim$ defined over all possible goods satisfying axioms 1 and 3. If G_1 and G_2 are thick goods which are weakly pointwise property equivalent and feature (perhaps) countably many property bundles, then A must be indifferent between G_1 and G_2.*

Proof. The only differences between this claim and that of theorem 1 are that (1) the goods G_1 and G_2 are not required to be bounded in space, and (2) the goods may involve a countably infinite number of property bundles. Let $p_1, p_2, \dots$ be an enumeration of the property bundles featured in both goods.

It can be shown [see Wagon 1985: 137] that any two subsets of $\mathbb{R}^3$ with nonempty interior are countably equidecomposible. Because both G_1 and G_2 are thick, it follows that the subgoods $G_1(p_i)$ and $G_2(p_i)$ occupy a subset of $\mathbb{R}^3$ with nonempty interior. Hence it is possible to find a countable decomposition of $G_1(p_i)$ which, when rearranged, will form $G_2(p_i)$. Let us denote the countable decomposition of $G_1(p_i)$ by $\{G_1^j(p_i)\}_{j=1}^{\infty}$. Note that Axioms 1 and 3 immediately yield the result that the agent A is indifferent between the subgoods $G_1(p_i)$ and $G_2(p_i)$, for all i.

Also note that this permits us to make a further fine-grain decomposition of the good G_1 into pieces in the following way:

$$\begin{aligned} G_1 &= G_1(p_1) \oplus G_1(p_2) \oplus \cdots \oplus G_1(p_n) \oplus \cdots \\ &= \left(\bigoplus_{j=1}^{\infty} G_1^j(p_1) \right) \oplus \left(\bigoplus_{j=1}^{\infty} G_1^j(p_2) \right) \oplus \cdots \oplus \left(\bigoplus_{j=1}^{\infty} G_1^j(p_n) \right) \oplus \cdots \end{aligned}$$

where the fine-grain decomposition involves only countably many pieces, since a countable sequence of countable sequences is itself countable. Applying Axiom 1 with the suitable isometries to the fine-grain decomposition then yields the result that A is indifferent between G_1 and G_2. ■

All of the theorems and proofs stated so far discuss conditions which force an ideal rational agent to be indifferent between two *single* goods G_1 and G_2, where one might naturally think of these goods along the lines of ordinary objects: goods occupying a single connected region of space. But there is no reason why the two goods *need* to be conceived of in this way. Given the proofs, it is clear that an ideal rational agent would have to be indifferent between, say, an alloy of gold and lead and a *collection* of two (separate) spheres of gold and lead, whatever their size. This observation enables us to generalize, with the aid of the following definition, the result of theorem 2 in a straightforward way.

Definition 6. Let $\mathcal{G}$ be a collection of goods. The *property bundle extension* of $\mathcal{G}$ is the set of all bundles p such that, for some good $\langle X, f \rangle \in \mathcal{G}$, there exists a point $x \in X$ such that $f(x) = p$. We denote the property bundle extension of $\mathcal{G}$ by $\| \mathcal{G} \|$.

Theorem 3. *Let A be an ideal agent who has a consistent and coherent preference ranking $\lesssim$ defined over all possible goods satisfying axioms 1 and 3. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are countable collections of thick goods such that $\| \mathcal{G}_1 \| = \| \mathcal{G}_2 \|$ and the cardinality of $\| \mathcal{G}_1 \|$ is countable, then A is indifferent between $\mathcal{G}_1$ and $\mathcal{G}_2$.*

Proof. Let $p_1, p_2, \ldots$ enumerate all the property bundles appearing in $\mathcal{G}_1$. Using the method of the previous proof, we know that the constraints placed upon A's preferences force A to be indifferent between the collection $\mathcal{G}_1$ and the collection $\mathcal{G}'$ consisting of nothing but pure goods, where the ith good is constituted out of the bundle p_i. But it is also true that A must be indifferent between the collection $\mathcal{G}'$ and the collection $\mathcal{G}_2$. Transitivity of preference means A is indifferent between $\mathcal{G}_1$ and $\mathcal{G}_2$. ■

4. What does it mean? (And responses to objections.)

The previous section provided several results which showed that an idealised rational agent whose preferences satisfy certain constraints must be indifferent between large families of goods. Let us take a moment to think through what this really means, and what it shows about the possibility of decision theory for idealised rational agents.

To begin, has it been shown that decision theory for ideal rational agents differs markedly from that of ordinary rational agents? In one important sense, it has. Recall the distinction Ullmann-Margalit and Morgenbesser propose regarding the difference between 'picking' and 'choosing':

> We speak of *choosing* among alternatives when the act of taking (doing) one of them is determined by the differences in one's preferences over them. When preferences are completely symmetrical, where one is strictly indifferent with regard to the alternatives, we shall refer to the act of taking (doing) one of them as an act of *picking*.
>
> [Ullmann-Margalit and Morgenbesser 1977: 757]

Now, although Ullmann-Margalit and Morgenbesser argue that we do, in fact, face real world 'picking' situations, I think it is fair to say that most of our decision theoretic problems involve *choosing* what to do. Yet what we have seen is that the decision theoretic problems concerning ideal rational agents, at least under one conception, mostly involve *picking*.

This difference matters. Recall the claim, made by Nover and Hájek, that decision theory 'is about rational decision makers in the abstract.' Part of the reason they stress this, I suspect, is because they believe we can learn important things about decision theory for ordinary people by taking seriously problems such as the Pasadena game, which only concern ideal rational agents. But can we *really* learn something about decision theory for ordinary rational agents by considering decision theory for ideal rational agents? The results from the previous section suggest not, for the two kinds of decision theory will have very different foci: *choosing* for ordinary rational agents, *picking* for ideal rational agents.

That said, it must be noted that although the idealised rational agent must be indifferent between absolutely enormously(!) large families of goods, the idealised rational agent is not, strictly speaking, indifferent between *everything*. This fact follows from the particular theory of objects we developed at the start of section 2.2. If objects were simply identified with spatiotemporal regions, then it would follow automatically[13] that any idealised rational agent whose preferences obeyed the assumptions would be indifferent between two objects occupying a region of space having nonempty interior. That would still not make the idealised rational agent indifferent between everything, but it would mean the agent is indifference between a larger class of goods than has been shown.

However, given the theory of objects developed, what do the indifference theorems show? Suppose that P is the set of all intrinsic properties; that is, P is the set of all properties which can be assigned to a point in space and time. (Hence the property 'is spatially extended' is not a member of P, nor is the property 'is between X and Y'.) The power set of P is the set of all possible property bundles which can be assigned to points of space. Any two regions of space having nonempty interior and assigned the same property bundle are equivalent under countable decomposition. Theorem 3 shows that any countable set of goods whose property bundles are assigned to regions of space having nonempty interior are equivalent — and thus the ideal rational agent is indifferent between those sets of goods. This means that the *only* things an ideal rational agent's preference relation concerns are the following:

Equivalence classes of objects having empty interior. None of the proofs in section 3 pertain to objects with an empty interior, so an ideal rational agent need not be indifferent between two objects where one has empty interior and another one which does not. Since two objects with empty interior may be equidecomposable under countable decomposition and rigid transformation, though,[14] the preference relation of an ideal rational agent must be taken to range over to *equivalence classes* of objects with empty interior.

Equivalence classes of objects having nonempty interior. As shown, the preference relation of an ideal rational agent has to range over equivalence classes of objects with nonempty interior because the Banach-Tarski theorem forces the agent to be indifferent between many of these objects.

Equivalence classes of nonmeasurable sets *and* sets of property bundles. The theorems of section 3 establish indifference results for sets of nonempty interior. However, some

[13] As before, from the version of the Banach-Tarski theorem involving decomposition into countable many pieces.

[14] Let G_1 denote the object consisting of points $x_i = (i, 1)$, for all nonnegative integers i that assigns property p_1 to x_i (where i is odd), and p_2 to x_i (where i is even). Let G_2 denote the object consisting of the points $x_{i'} = (i, 2)$ that assigns property p_1 to x_i (where $i = 1,4,7,10, ...$) and property p_2 to all other x_i. It's trivial to see that G_1 and G_2 are equidecomposible under countable decomposition and rigid transformation by the type of trick used to rearrange the guests in Hilbert's Hotel. Hence an ideal rational agent will be indifferent between G_1 and G_2.

nonmeasurable sets are equidecomposable under countable decomposition and rigid transformation. Given two objects (these are highly idealised objects!) which happen to be of this type, and whose decomposed parts have compatible assignments of property bundles, an idealised rational agent will be indifferent between them. This indifference relation generates equivalence classes. An ideal rational agent's preference ordering, however, has considerable freedom in determining an ordering over these equivalence classes.

Potential trade-offs. This is messy. Notice, firstly, that an ideal rational agent will be indifferent between (1) an equivalence class of objects having nonempty interior and (2) an equivalence class of nonmeasurable sets and sets of property bundles. Why? Consider the proof of theorem 1. When we took the thick good G_1 and cut it into a finite number of pieces $\mathcal{P}$ so as to rearrange and form G_2, each of the pieces of $\mathcal{P}$ was a nonmeasurable set. The ideal rational agent was thus indifferent between G_1 and the collection $\mathcal{P}$. Hence the agent would also be indifferent between G_1 and the equivalence class defined by $\mathcal{P}$. And *that* means the ideal rational agent would be indifferent between the equivalence class containing G_1 and the equivalence class containing $\mathcal{P}$.

Nevertheless, it will, presumably, be the case that there exist equivalence classes of nonmeasurable sets which the ideal rational agent does *not* treat as equivalent to some thick object. In this case, the preference relation of the ideal agent will have to specify which of the two equivalence classes should be preferred. This is the nature of the 'potential trade-offs'.

Having said that those are the *only* things an ideal rational agent's preference relation may concern sounds like a droll understatement. Those equivalence classes provide quite a number of items for the preference relation to sort! Yet notice that in the world of the ideal rational agent, the importance of *measure* largely disappears: all that matters about a good is whether it has empty or nonempty interior. Once two goods share their property bundles on a set of nonempty interior, they are equivalent.[15] The sense, then, one can speak of ideal rational agents *choosing* is a notion of choice for which measure or quantity is largely irrelevant. Ideal decision theory looks a good deal different from ordinary decision theory where we do often care about the measure and quantity of goods.

4.1. Defending the assumptions.

The first place to attack any argument is with the assumptions, so let me offer a pre-emptive defense against four objections. First, that the conception of a 'good' adopted here deviates from one frequently used in decision theory: goods (or outcomes) are thought of as a set of fully specified possible worlds.[16] Second, the way goods are defined omits relational properties which are often relevant for choice. Third — but related to the first point — not all goods are such that such that a person's preference relation is invariant under rigid transformations. And, finally, goods in the real world are often 'lumpy' (that is, not infinitely sub-dividable). Since the proofs in section 3 require all of these assumptions to go through, rejecting any one suffices for blocking the result.

Regarding the first objection, let me acknowledge that a tradition exists of treating the preferences of an agent as ranging over fully specified possible worlds, but that I also think there is good reason for deviating that tradition. Treating outcomes as sets of fully specified possible

[15] Here it is important to note that objects, on our definition, are taken to exist within $\mathbb{R}^3$. Because the *interior* of a set is the union of all its open subsets, a object with nonempty interior contains at least one open subset, and hence has positive measure. The other direction need not hold, though: there are sets with empty interior having positive measure [Csörnyi et al. 2006].

[16] I would like to thank an anonymous referee for calling to my attention the fact this was not discussed in an earlier version of the paper.

worlds undoubtedly has certain advantages from the point of view of developing elegant formalisations. However I think that conceiving of outcomes in this way proves to be a disadvantage when performing conceptual analysis on the ideas of choice, preference, rationality, and the rationalisation of choice.

Consider the following: suppose an agent must choose between a bar of gold or a bar of lead. From the point of view of fully specified possible worlds, there is nothing irrational or incoherent about an agent with the following set of preferences:

$$
\begin{aligned}
\text{gold bar} &\gtrsim \text{lead bar} \gtrsim \text{gold bar} - 1\text{ atom} \\
&\gtrsim \text{lead bar} - 1\text{ atom} \gtrsim \text{lead bar} + 2\text{ atoms} \\
&\gtrsim \text{gold bar} + 2\text{ atoms}
\end{aligned}
$$

and so on. From the point of view of *fully specified possible worlds*, these are all different, and so there's no way of being able to ground a charge of irrationality or incoherence.

I suggest that there is — or, rather, there should be — something which strikes us as deeply disconcerting in taking this as a coherent preference order. What? Namely that if we were to encounter another human whose preferences took this form (imagine that preference ordering being extended in the same bizarre way over everything) we would be incapable of understanding that person as having coherent preferences. Why? Because when faced with that pattern of choices (imagine that they are revealed through choice behaviour) we would be unable to *rationalise* the preferences of the agent to ourselves. We would be unable to think our way into what motivated the agent to choose *like that*.

The point is that when we talk about coherent preferences among real agents, we mean something strictly stronger than just a coherent preference ordering over fully specified possible worlds. We need to be able to rationalise the agent's choice behaviour to ourselves. Rationalising choice involves describing the agent's choice in terms of choosing types or kinds of goods, as in 'she prefers gold jewelry over lead jewelry, unless the lead jewelry is phenomenally well crafted'. When we cannot provide such a rationalisation, that often gives us strong evidence for attributing irrationality to the agent. But treating outcomes as fully specified possible worlds, and thereby concentrating on the maximally possible fine-grain distinctions that could be made, largely deprives us of the ability to attribute irrationality to agents. There is almost always *some* minor difference that exists which *could* support a coherent preference order.

And *that* is the reason why I think it is useful, on occasion (such as here), to treat goods as components, or parts, of worlds. The preferences of real agents range over parts of worlds. My own preferences are largely unaffected by the wide variety of Cambridge changes which constantly alter the properties of the objects around me. Hence, it is worth thinking about what kinds of changes to goods would, for a rational agent, not alter his or her preferences. The axioms proposed in section 3 provide one such set of suggestions for an ideal rational agent.

The second objection concerning relational properties appears plausible on first blush. Suppose that two tract houses, identical in construction and lot size, are for sale. The first house overlooks a disused petrol station in Slough, the second overlooks the English channel off Chesil Beach. The fact a rational agent should be willing to price these two houses differently, one might argue, implies that it is not just the *intrinsic* properties of a good which matter, but *relational* properties as well.

Yes, one could argue that. But one could equally well argue that what the choice concerns are two goods with greater spatial extent than normally conceived.[17] Any choice problem which initially appears to involve some good G with relational properties $p_1, \dots, p_n$ can be equally well framed as a choice problem involving some good G' with only intrinsic properties. How?

[17] On this view, the legal rights you acquire upon purchasing the house and land only concerns a subset of the actual good.

Let $O(i)$ denote the object (or set of objects) upon which the relational property p_i depends. Then let G' be the *mereological sum* of G and all the $O(i)$. All the relevant relational properties of G become intrinsic properties of G', but the choice problem is unchanged.

What of the third objection, that not all goods are such that a person's preference relation is invariant under rigid transformations? What motivates this worry, I take it, is something like the following: suppose I offer you a cup of coffee, for which you would be willing to pay \$1. Now suppose that, at the last moment, I turn the cup of coffee upside down. Should you still be willing to pay \$1? This objection is related to the concern about relational properties as the following variant of the house-purchase example illustrates: it seems that the fact you would be willing to pay X for a house on Chesil Beach does not imply you would be willing to pay X for the same house if I were to jack it up and move it to Mudchute. Don't both of these examples contradict the claim that our preference ordering of goods should be invariant under rigid transformations?

No. To begin, notice that the claim about rigid transformations only concerns whether your preference ordering should be affected by the mere fact that a rigid transformation *has been performed*. It says nothing about whether your preference ordering should be affected if, after the transformation has been performed, we start the clock and let the ordinary causal forces in the universe unfold.[18] It is this latter worry which drives the coffee cup example, for obviously once the causal clock starts ticking the coffee will spill out, destroying the good.

Once we note that the future causal consequences of rigid transformations are excluded from the question of whether a transformation should influence your preference ordering, it follows that it is, in fact, rational to require preference invariance under such transformations. Here's an extended thought experiment to illustrate why: We all know that it is possible to fill a bucket with water and rotate it via the handle so that the bucket will, for an instant, be upside down without spilling a drop. Now imagine doing the same thing but with the cup of coffee at the bottom of the bucket. Suppose when the bucket reaches its apogee I offer to sell you the cup of coffee (which, through the action of centripetal force is still nicely situated at the bottom of the bucket). If you were willing to pay \$1 for the cup of coffee right-way-up, you should be willing to pay \$1 for the cup of coffee upside down. If not, you create arbitrage opportunities for a talented bucket spinner to money-pump you.[19]

Furthermore, notice that the solution to the objection about relational properties also

[18] If so, then no preference ordering would be invariant under rigid transformations, even for objects for which it makes sense, like cans of beer: simply ask the question about what your preference would be after transforming it to the middle of the sun.

That said, let me both acknowledge a potential worry and explain why I am not treating it at further length. I grant that decision theory often does take into account future causal consequences. Yet although decision theory often does this, it does not always do so. More importantly, any discussion of how future causal consequences affect choice will need to invoke such things as the individual's discount rate and the extent to which the individual can anticipate or predict future preference shifts he or she might have. For example, if I go to a fancy restaurant where I need to order an expensive and laborously prepared dessert at the beginning of the meal, I need to try to determine before the meal what my post-dinner preferences will be. That may be difficult. I might want a vanilla mousse with whipped cream *now*, but after consuming an acidic tomato-and-garlic dish, that might be the last thing I want later. Both of these matters are difficult ones to address, even when dealing with ordinary rational agents for which they most naturally make sense (i.e., discount rates due to finite lifespans, and so on). Do ideal rational agents have discount rates? If so, why? What would they be? Thus I set aside these issues, even though I appreciate that they may be causes of concern, simply because they fall outside the main focus of this paper.

[19] Suppose you value the upside-down coffee for some amount $X > 1$. If you originally own the coffee, the bucket-spinner can buy it off you for \$1, put it in the bucket and rotate it. At the apogee, he can sell it to you for $\$X$. Once you are holding the coffee again, he can repeat the process and make a profit of $\$(X - 1)$ each iteration.

Now suppose you value the upside-down coffee for $X < 1$. If the bucket-spinner originally owns the coffee, he can sell it to you for \$1. If he then puts it in the bucket and rotates it, at the apogee he can buy it off of you for $\$X$. Once he stops spinning the bucket, he can sell it to you for \$1, again. Each iteration the bucket-spinner makes a profit of $\$(1 - X)$.

handles the variant of the house-purchase example. The reason why jacking up the house on Chesil beach and moving it to Mudchute changes the value is because the good to which the price attaches *isn't just the house*. The good being priced is the mereological sum of the house and, well, the English channel, plus more. Moving the house to Mudchute changes the nature of the internal relations between some of the proper parts, yielding a different mereological sum.[20]

Now consider the fourth objection: goods in the real world are lumpy and not infinitely sub-dividable. This attacks the use of the Banach-Tarski theorem in the proofs, for showing any two sets with nonempty interior are equidecomposible into countably many pieces requires taking one set apart into countably many nonmeasurable point sets. Lumpy goods are not continuously divisible, and so the construction cannot be completed.

Although it is true that many real world goods are more like telephones than pies, this objection basically misses the point. In a world in which the axiom of choice is true, even a lumpy good could be cloned via the Banach-Tarski construction. It's just that the pieces of the lumpy good, during the intermediary stages of the construction, might be valued for less than the good(s) at the beginning and final stages of the construction.

There are several issues lurking here that one might take umbrage with. To begin, one might object that although the ideal rational agents of decision theory know their own minds and know all logical and mathematical truths, they do not have any extraordinary abilities to manipulate goods. In particular, they do not have the ability to manipulate goods in the way required by the Banach-Tarski theorem.

That’s a fair enough worry, but notice that neither axiom 1 nor 2 (nor, for that matter, axiom 3) assume that ideal rational agents have any special ability to manipulate goods whatsoever. Those axioms only require that the ideal rational agent's preference relation remain constant under certain hypothetical, or potential, transformations. Those transformations need never actually *be performed*; and even if they were, the actual actor responsible for doing so does not matter. The point being that if an ideal rational agent’s preferences were, in fact, constant under those hypothetical transformations, then an ideal rational agent would be indifferent between two balls of gold regardless of their size or quantity, even though the Banach-Tarski decomposition and recomposition *had never been performed*.

A further issue one might have is that all transformations are treated equally: transforming a good one centimetre to the right has the same effect — namely, none — on the ideal rational agent's preferences as transforming the good ten thousand metres to the right. Real human agents, so the objection goes, would respond very differently to these cases. The reason why some transformations do not affect a person’s preferences is because they are either costless to undo or do not alter the members of the feasible set (i.e., shifting a good one centimetre to the right). When transformation is such that the cost to undo becomes excessively high, or that it effectively removes the good from the feasible set (i.e., shifting a good ten thousand metres to the right), the transform would alter the person’s preferences. Given this, the objection continues, the failure to allow an ideal agent’s preferences to be affected by the cost of the transform is implausible.

Now it is undoubtedly true that real people’s preferences may be influenced by the cost of transformations. But real people’s preferences may be influenced by the cost of solving logical and mathematical problems, or of working through the implications of their own beliefs. These costs are readily ignored by decision theorists in their creation of the ideal rational agent, even though they, too, hugely influence the behaviour of real people. Given this, I fail to understand the ground rules for creating these Galilean idealisations. Why is my ignoring the cost of transformations considered problematic whereas their ignoring the cost of logical and mathematical inference is not? (To say nothing of the fact that the ideal rational agent is assumed to know all mathematical truths, even those truths which are not provable!)

[20] Note that the conception of mereology invoked here differs from that of classical extensional mereology [Casati and Varzi 1999] in that it denies that any two objects with the same proper parts are identical (the internal relations between parts matters, too).

But doing so misses sight of the purpose of the exercise: to consider whether it made sense to speak of decision theory for ideal rational agents. Or, as Sydney Morgenbesser might have put it: We are talking about ideal rational agents who have *unbounded utility functions*, who can solve *arbitrarily complicated* problems of rational inference in a flash, and who have a *complete and coherent* preference ordering over *all possible goods*, and *you're worried* about whether transformations should have costs?

4.2. The Axiom of Choice.

It is, I suppose, rather fitting that the problems which the Banach-Tarski paradox generates for ideal decision theory derive entirely from the axiom of choice.[21] If we are interested in formulating a decision theory 'not merely about humans' but also about 'ideal rational agents' (as Nover and Hájek suggest), it is hard to see how we could *not* include, at least potentially, the axiom of choice among our assumptions. Given an interest in ideal rational agents, it is a perfectly *natural* assumption to adopt.

Why do I say that Choice is a natural assumption to adopt? Recall the standard formulation of the axiom [Jech 1978, 1]:

Axiom of Choice: Every family of nonempty sets has a choice function.

More precisely, if $\mathcal{F}$ is a family of nonempty sets, then there exists a function f such that for every $X \in \mathcal{F}$, $f(X) \in X$. One way to think about this is in terms of an ideal rational agent always being able to work through a family of nonempty sets, selecting one item from each set. Indeed, this was the image invoked by Bertrand Russell in his famous anecdote that, given an infinite pile of pairs of shoes and socks, one needs to invoke the axiom of choice to form a set created by selecting one sock from each pair.

Now, I freely admit that being able to perform such a hypertask isn't *necessarily* constitutive of the concept of an 'ideal rational agent'. We can certainly formulate idealisations of real human agents in other ways, where the idealised agent's choice powers are understood differently from possessing the full power of the axiom of choice. But consider the following argument: suppose that the choice problem simply is the entire set-theoretic universe $\mathbb{V}$. The ideal agent's preferences over sets thus corresponds to an ordering of all the sets.[22] One reason why we might consider an ideal agent to possess the full discriminatory powers of the axiom of choice is that it provides an idealisation of revealed preference: given a family of nonempty sets, each member of the family is presented to the ideal agent, who selects an item from it. At the end of this process, the set of objects the ideal agent has is one corresponding to a choice function on the original family of sets.

In any case, what we see is that there are a number of ways of theorising about ideal rational agents, each of them with differing capabilities of choice. Instead of full choice, there are a number of possible substitutes. For example, we could allow the axiom of countable choice, or the axiom of dependent choice: [Jech 1978, 41]:

Axiom of Countable Choice: Every countable family of nonempty sets has a choice function.

Axiom of Dependent Choice: If R is a binary relation over a nonempty set A, and if for every $a \in A$ there exists a $b \in A$ such that bRa, then there is a sequence $a_0, a_1, \dots, a_n, \dots$ in A such that $a_{n+1}Ra_n$.

[21] Recall that in section 3.1 I assumed that an ideal rational agent believed some extension of ZFC.

[22] Is this possible? Yes, under some conceptions of set theory. In his proof that the generalized continuum hypothesis is consistent with ZFC, Gödel invoked the axiom of constructability. The axiom of constructability implies the axiom of choice as well as the possibility of well-ordering the set theoretic universe.

I mention both axioms because, although the restriction to countable choice might first come to mind, dependent choice turns out to be stronger than countable choice (although clearly weaker than the full axiom of choice), and actually implies it.

The reason why this matters is that, if we consider a decision theory which aims to describe ideal rational agents capable of dependent choices (rather than the full axiom of choice), the proofs of section 3 *no longer go through*. One can show [Wagon 1985: 207–210] that the Banach-Tarski theorem cannot be proven in $\mathrm{ZF} + \mathrm{DC}$. Why? It turns out that $\mathrm{ZF} + \mathrm{DC}$ is consistent with the claim that all subsets of the $\mathbb{R}^3$ are Lebesgue measurable. Since the Banach-Tarski paradox decomposes the sphere into nonmeasurable subsets, if no nonmeasurable sets exist, then the Banach-Tarski paradox cannot hold. And, hence, an ideal rational agent whose preference ordering remains invariant under countable decomposition and rigid transformation is not forced to be indifferent across entire families of goods.

Yet this simply underscores the point which I have been stressing throughout: it is unclear just what, exactly, we mean by the phrase 'ideal rational agent'. One arguably natural precisification yields an ideal decision theory having little relevance for ordinary decision theoretic problems. Alternative idealisations may, of course, imply different claims. Perhaps we might take the axiom of determinacy to underwrite the capabilities of ideal rational agents.[23] The axiom of determinancy contradicts the axiom of choice, so again the proofs of section 3 are blocked. Yet why is determinancy a better axiom to adopt than choice? Not all real-world games have winning strategies, and so it remains unclear why an idealisation based on winning strategies for a class of two-player games appears more sensible than an idealisation based on revealed preference.

5. Conclusion.

As we strive to formulate theories of how rational agents ought to make decisions and play games, it is natural to appeal to some notion of an 'ideal rational agent'. After all, real human agents, with their limited problem solving ability and tendency to make irrational choices, make formulating a *descriptively* accurate decision theory extremely difficult. (To say nothing of the challenges of formulating a normative decision theory.) Yet one must be careful in how one conceives of this 'ideal rational agent'. As we have seen, an ideal rational agent whose preferences and beliefs satisfy the assumptions of section 3 winds up being forced to be *indifferent* among entire families of goods. This indifference means that decision theory for ideal rational agents looks significantly different from decision theory for finitary humans, and hence serves to shed little light on the subject which is our ultimate concern.

It would be wrong to say that section 3 provides an *impossibility* result regarding a decision theory of suitably idealised rational agents. But it is fair to say that the results establish an upper limit on how far we can idealise our conception of a rational agent before the very concept of choice starts to become inapplicable. It is, you might say, a *comprehensibility* result: you can idealise *this* far, but no further, as the concepts which you are seeking to elucidate begin to fall apart.

But if we step back from the edge of idealisation, we find that considerable space remains to be explored. In the world of $\mathrm{ZF} + \mathrm{DC}$, or of $\mathrm{ZF} + \mathrm{Determinacy}$, we find other conceptions of an ideal rational agent who must no longer be indifferent among entire families of goods. What more can be said remains to be discovered.

[23] The axiom of determinacy states that the following family of two-player games has a winning strategy. Let S be the set of all infinite sequences of natural numbers, and let $A \subset S$. The game is played as follows: player I chooses a natural number a_0, then player II chooses a natural number a_1, and players I and II continue to alternate their choices in this fashion so as to construct the sequence $\sigma = \langle a_0, a_1, a_2, a_3, \ldots \rangle$. Player I wins if $\sigma \in A$; otherwise, player II wins.

Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science

References

Arora, Sanjeev and Boaz Barak 2009. *Computational complexity: A modern approach.* Cambridge University Press, 2009.

Casati, R. and A. Varzi 1999. *Parts and Places: The Structures of Spatial Representation.* MIT Press.

Csörnyei, M., T. Jordan, M. Pollicott, D. Preiss and B. Solomyak 2006. Positive-measure self-similar sets without interior. *Ergodic Theory and Dynamical Systems* 26: 755–758.

Davidson, Donald 1974. On the Very Idea of a Conceptual Scheme. *Proceedings and Address of the American Philosophical Association,* 47: 5–20.

Easwaran, Kenny 2008. Strong and weak expectations. *Mind* 117/467: 633–641.

Elster, Jon 1985. The Nature and Scope of Rational-Choice Explanation. In E. LePore and B. McLaughlin, editors, *Actions and Events: Perspectives on Donald Davidson*, Blackwell Publishers: 60–72.

Fine, Terrence L. 2008. Evaluating the Pasadena, Altadena, and St. Petersburg Gambles. *Mind* 117/467: 613–632.

Jech, Thomas 1978. *Set Theory*. Academic Press, Inc..

Nover, Harris and Alan Hájek 2004. Vexing expectations. *Mind* 113/450: 237–249.

Pettit, Philip 1995. The Virtual Reality of *Homo Economicus. Monist* 78: 308–329.

Thaler, Richard H. and Cass R. Sunstein 2009. *Nudge: Improving Decisions about Health, Wealth and Happiness*. Penguin.

Ullmann-Margalit, Edna and Sidney Morgenbesser 1977. Picking and Choosing. *Social Research* 44/4: 757–785.

Wagon, Stan 1985. *The Banach-Tarski Paradox*. Cambridge University Press.