Rational theory choice: Arrow undermined, Kuhn vindicated

Seamus Bradley

December 1, 2013

In a recent paper, Samir Okasha presented an argument that suggests that there is no rational way to choose among scientific theories. This would seriously undermine the view that science is a rational entreprise. In this paper I show how a suitably nuanced view of what scientific rationality requires allows us to avoid Okasha's conclusion. I go on to argue that making further assumptions about the space of possible scientific theories allows us to make scientific rationality more contentful. I then show how such a view of scientific rationality fits with what Thomas Kuhn thought.

1. Introduction

Scientists are often faced with many competing theories. How are they to choose among them? Thomas Kuhn suggested that there are various virtues that theories might have that recommend them to us (Kuhn [1977] 1998). So one theory might fit very well all the data we have. Another theory might not fit as well, but be more simple or elegant. One theory might have the virtue of being very fruitful. How are we to choose when the theoretical virtues are pulling in different directions? Kuhn argued that there needn't be a right answer to this question. That is, many different algorithms for trading one virtue off against another may be legitimate theory choice rules (Kuhn [1962] 1992). Call this his "many algorithm" claim.

Recently, Samir Okasha has suggested that it might be that there is in fact *no* rational theory choice algorithm (Okasha 2011). He does this by an ingenious argument that imports Kenneth Arrow's famous "impossibility theorem" from social choice theory into the context of scientific theory choice.

My aim in this paper is to argue that with an appropriately nuanced understanding of what rationality requires of us in the context of theory choice, we can escape Arrow's impossibility. Furthermore, a secondary goal of Okasha's paper was to "illustrate. . . how techniques from theoretical economics can be applied to problems in epistemology" (p. 84) and I show how such input can add content to the claim that science is a rational entreprise.

The outline of the paper is as follows. I first briefly reconstruct Okasha's argument, and summarise Arrow's impossibility result. Second, I argue that we need a subtle view of what rationality requires. With this nuanced understanding of rationality, I return to the problem of theory choice and show that theory choice is still rationally constrained, even if it is not rationally determined. I then suggest that this strategy is somewhat implicit in Kuhn's own remarks on the subject.

2. Okasha's argument

Let's imagine that we have some collection of theories $\{\alpha, \beta, \dots\}$ and we are trying to choose between them. There are various theoretical virtues that these theories have to varying degrees. Let's say that each virtue determines a linear order R on the theories. That is, R is a relation on the set of theories that is complete and transitive. Now, let's imagine that α is simpler than β, but β is more accurate. Then if R_s and R_a are the relations of "is more simple than" and "is more accurate than" respectively, then $\alpha R_s \beta$ but $\beta R_a \alpha$.

What we want is a theory choice rule. This is a rule that tells us how we should trade off more simplicity versus more accuracy and so on. In short, it is a function that takes the various relations associated with the theoretical virtues as inputs, and outputs a single relation which is the aggregate "is a better theory than" relation. If $\mathcal{R}$ is the set of all complete transitive relations that can be defined on our set of theories, then a theory choice function is a map from $\mathcal{R}^n$ to $\mathcal{R}$ where n is the number of virtues we are concerned with. That is, the function takes a *profile* of orderings which reflects how each theoretical virtue orders the theories and it outputs another ordering which is the aggregate goodness order on the theories.

We want our function to have a *Universal Domain*: that is, we want our function to be able to cope with any possible profile of orderings over the theories. What other properties might we want a rational theory choice rule – or theory ordering rule – to satisfy?

It seems that if theory α is better than β according to all of the theoretical virtues, then this unanimity should be reflected in the aggregate ordering. This is known as the *Pareto Condition* and seems like another thing we should require of our theory choice rule.

We have several theoretical virtues and we should take them all into account at least to some degree. It would be problematic if there were one particular virtue whose ordering judgements were always reflected in the aggregate relation. This would mean that the other theoretical virtues are not really having an effect on theory choice.[1] We don't want theory choice to be dominated by a single virtue in this way. So we should require that our theory choice rule satisfy a condition of *Non-dictatorship*: no virtue should be a "dictator" in the sense of *always* dictating what the aggregate relation is like.

[1] There's a subtlety here that I am going to gloss over. See Okasha's remarks on lexicographic orderings for the details (Okasha 2011, pp. 95–6).

Finally, the aggregate relationship between α and β should depend *only* on the individual virtue's ordering of α and β. That is, how the virtues rank other theories should not affect the aggregate ranking of α and β. This condition is known as *Independence of Irrelevant Alternatives.* IIA is probably the most controversial of the conditions and my cashing out of it here doesn't adequately reflect the way in which it is controversial or problematic.

So we would like a theory choice rule that has a *Universal Domain* and satisfies *Pareto Condition*, *Non-dictatorship* and *Independence of Irrelevant Alternatives.* Sadly, as Kenneth Arrow showed, there is no such rule.[2] That is, no function satisfies those four properties. So if we take all of these properties as requirements on rational theory choice, it looks like there is *no rational theory choice algorithm.* This is Okasha's surprising conclusion.

Let's compare this conclusion with Kuhn's claim. First, what does Kuhn's original "many algorithm" claim mean for scientific rationality?

> This does *not* mean that theory choice is irrational, Kuhn stressed, or that 'anything goes', but rather that the traditional conception of rationality is too demanding... [T]he '[many] algorithm' argument does not undermine the rationality of science, he thinks, but rather forces us to a more realistic conception of what rational theory choice is like. (Okasha 2011, p. 86)

Later Okasha argues that one response to Kuhn's "many algorithm" claim is to

> liberalize the notion of rationality, and argue that two scientists could both count as rational despite employing different algorithms for theory choice. (p. 94)

In later sections I will do exactly that.

Okasha's "no algorithm" claim is much stronger. While Kuhn's view leaves open the possibility that further constraints might determine rational theory choice; if Okasha is right, no such possibility exists. If the Arrow-inspired impossibility is correct, then no theory choice algorithm satisfies even the minimal constraints on rational theory choice. This would be a serious blow to scientific rationality. If many rational theory choice algorithms exist, then theory choice is (at least partially) rational. If no such algorithm is possible, then theory choice is fundamentally irrational: however scientists choose among theories violates some principle of rationality. My aim is to give a particular gloss on how to liberalize scientific rationality and show how Arrow's impossibility result, once suitably defanged points the way to a view of theory choice consonant with Kuhn's view.

3. Rationality can be silent

In this section I want to urge a nuanced view of what rationality requires. The aim is to use this view in the next section to escape from Okasha's negative conclusion. Let's

[2] Arrow's theorem was originally presented in the context of voting. The relations are interpreted as individuals' preferences and the aggregate relation is the output of a voting rule. Gaertner (2009) gives three different proofs of Arrow's theorem.

start by considering a simple example of rational choice. A fair coin is about to be tossed. You have the choice to bet on heads or on tails at the same odds. Which should you choose? Both bets have the same expected value, so the standard rational decision theory apparatus doesn't discriminate between the bets. Both bets are equally good by the lights of your epistemic state, so there is nothing to choose between them: rationality gives you no advice as to which to choose. So, *rationality can be silent.* But note that just because which bet to take isn't rationally *determined*, that does not entail that choice in such a betting scenario is not rationally *constrained.* It would still be irrational to choose a bet on heads at shorter odds if a bet on heads at longer odds were available.

Imagine you are given the choice between two totally incommensurable goods A and B. Rationality is silent as to which of A and B you should take. That's not to say that anything goes: if there is a third good $B+$ which is strictly better than B, then you should prefer $B+$ to B. So while rationality is silent on some questions, it still constrains choice. Importantly, in a choice between A and B, it would not be irrational to choose A.

You are offered the choice between infinitely many options of the form "n days in heaven followed by an eternity in hell". Arguably, no choice of n is rational since there is always a bigger n. This seems to be a decision problem with no rationally sanctioned resolution.[3]

There are two boxes on the table. One contains £100, the other contains nothing. You get to choose a box, on the understanding that you will be given £1000 if and only if you act irrationally.[4] Given this understanding, acting irrationally earns you more money and is therefore the rational action. One might respond to this by arguing that rationality fails to give you any advice in this example. No action can be rationally justified.

These last two cases suggest that in some circumstances, rationality appears to give you no useful advice because the problem set-up makes it impossible to choose rationally: every choice is irrational. The constraints rationality imposes on choice cannot be satisfied because of the structure of the objects of choice. This is in contrast to the first two cases where rationality was *silent* on some kinds of questions, but still constrained choice in certain ways: there were still certain choices that were not irrational. The constraints on choice can be satisfied, but such satisfaction does not necessarily determine a choice.

The above examples suggest that there are two ways rationality can fail to give you advice: rational silence and incompatible constraints. The first two examples highlight the first kind of case; the last two examples, the second. Arrow's theorem is a case of incompatible constraints. I argue that the rationality involved in rational theory choice is actually a case of rational silence. Theory choice can be rationally constrained without being rationally determined. We should make room for rational silence in theory choice. Once I've articulated what this means in more detail, I will show that Kuhn's position in Kuhn ([1977] 1998) is very close to this.

[3] Thanks to Jim Joyce for pointing out this example to me.

[4] The idea of this sort of game is from Gaifman (1983).

4. Arrow undermined

To recap, Arrow's imposibility theorem says that there is no function that takes a profile of individual orderings and outputs an aggregate ordering that satisfies the four conditions listed above. It is important to notice that what the theorem rules out is a function that outputs an *ordering* of the theories: a complete and transitive relation on the theories. That is, the codomain of the aggregation function is $\mathcal{R}$: the collection of linear orders. What if we asked, instead, only for a partial ordering of the theories? That is, what if we didn't demand that the aggregate goodness relation on theories be complete? Might such a theory choice rule be possible? The answer is yes.

Any satisfactory theory choice rule must satisfy the Pareto condition. This means that when all the theoretical values agree, the aggregate relation must respect this unanimity. This means that the aggregate relation is a superset of the intersection of all the individual relations. That is, we can think of a relation R as a collection of ordered pairs $[R]$ of its domain where $(\alpha, \beta) \in [R]$ just in case $\alpha R \beta$. So it makes sense to think of subsets and supersets of relations, as well as intersections and unions of relations. Note that the intersection of all the individual relations is just the collection of pairs that are in every relation, which is just the pairs that have to be in the aggregate relation in order for it to satisfy the Pareto condition.

Let's think about this relation, let's call it $R_\cap$. That is, $\alpha R_\cap \beta$ just in case $\alpha R_i \beta$ for all theoretical virtues i. It is transitive: this follows from the transitivity of the individual relations. It is not likely to be complete: all that it takes for $R_\cap$ to fail to be complete is that there are relations R and R' among those being aggregated, and theories α and β such that: $\alpha R \beta$ but $\beta R' \alpha$. In such a case, $R_\cap$ will hold in neither direction between α and β. $R_\cap$ could quite possibly be empty. Consider the case above where α and β are the only theories. Then $R_\cap$ imposes no constraint on theory choice at all. When $R_\cap$ is not empty, however, it imposes a reasonable constraint on what counts as rational theory choice. Such a relation can be generated for any profile of orderings so the function that returns this intersection relation has universal domain. It will also satisfy Non-dictatorship and IIA.[5]

If we were to adopt the function that returns this relation as our theory choice rule, we would often get very little advice. When we did get advice it would be good advice, but often the relation would not hold either way between two theories. The relation would not help us make a *choice* between those options. On the standard (strong) understanding of rationality, this failure to determine a best theory would be seen as a pretty major flaw. I would like to suggest instead that this rational silence is a good thing: it could be an indication that both theories are legitimate objects of study given the current state of evidence.

Often a rule like $R_\cap$ will not determine theory choice. That is, there will be several theories which are "at the top" of the aggregate ordering, but that are incommensurable in the sense that $R_\cap$ doesn't hold between any pair of the theories. That is, there will

[5] No R can be such that for any profile including R, if $\alpha R \beta$, $\alpha R_\cap \beta$, since there is always a profile including an R' with $\beta R' \alpha$ which means that for that profile, $R_\cap$ holds in neither direction. It is obvious from the way that $R_\cap$ is constructed that IIA will also hold.

be cases where theories α, β will have the property that there is no γ such that $\gamma R_\cap \alpha$ respectively β, but $R_\cap$ does not hold between α and β in either direction. In such a situation the theory choice rule has failed to determine a choice. Rather than seeing this as a failure of rationality, I think we should see this as a legitimate case of rational silence. As we will see later, rational silence is not just legitimate, but it is in fact a *desirable* property of scientific rationality. Rationality is often considered to be that which determines choice; I am arguing that rationality should instead be thought of as something that merely *constrains* choice.

Arguably however, the theory choice rule $R_\cap$ is not nearly discriminating enough. That is, one can imagine a situation where one theory (α) is only marginally simpler than another (β), but β is so much better fitted that it would be crazy to stick to the marginally simpler theory α. However, $R_\cap$ would be silent on these two theories, since one is simpler and the other is better fitted and thus there is no unanimity as to which is better. I will have a lot more to say about that in the next three sections, but first I want to point out that even though $R_\cap$ is clearly not all there is to theory choice, it is enough for a certain purpose: it shows that Arrow's theorem has been undermined. That is, it shows that if we allow our aggregation rule to output a merely *partially ordered* relation, then there do exist rules that satisfy all of Arrow's criteria. In this sense, Arrow's theorem has been undermined. I find such an undermining of Arrow's theorem somewhat unsatisfactory and the next three sections are an attempt to go beyond this unsatisfying resolution of the problem.

Before moving on, I would like to emphasise that $R_\cap$ captures some kind of minimal constraint on rational theory choice. That is, however preference among theories is determined, it should be a superset of $R_\cap$. So it is not the case that anything goes! There are kinds of aggregate preference among theories that are ruled out as irrational. The constraint is very minimal, but it is a constraint nonetheless. Without making further assumptions about the structure of the space of possible scientific theories, this is, perhaps all we can say. In terms of the constraints on rational theory choice in the abstract, this is perhaps all we can say. This at least shows that we are in a rational silence regime, not an incompatible constraints regime. If I thought that that was all there was to say about scientific rationality, I would end the paper here. I don't think $R_\cap$ exhausts what we can say about scientific rationality: I think scientists' choices among theories are more constrained than simply avoiding "dominated" theories. The nature of these further constraints will be outlined in the next few sections.

5. The informational basis escape

After having argued that Arrow's theorem as applied to theory choice shows – contra Kuhn – that there is *no* rational theory choice algorithm, Okasha suggests a possible escape route for theory choice. Arrow's theorem works on the assumption that the individual virtues that are aggregated are merely *ordinal.* That is, one can only say that α is simpler than β, not *how much simpler* α is. Following the work of Amartya Sen, Okasha shows that theory choice is possible under certain assumptions. The details don't matter for the moment so I am going to sweep a lot of complexity under the carpet by describing the assumptions in the following sloppy way:

- The individual virtues provide cardinal information, not just ordinal information. This means roughly that there is a real-valued function (unique up to affine transformation) on the theories that represents how much of a certain virtue a theory has. The difference between α's simplicity and β's simplicity is meaningful.
- There is some rate of exchange that tells you how much of one virtue is worth trading off against how much of another virtue. This means, roughly, that if α is simpler and β more accurate, you know how much more accurate α would have to get in order to be as good as β.

So, very roughly, if the aggregation takes account of *how much* of each virtue a theory has, and if the aggregation can meaningfully trade off one virtue against another, then aggregation is possible. Thus, this provides a way out of the impossibility. Note that this escape route amounts to denying IIA (see section 6 of Okasha (2011) for details). Okasha thinks that such an escape route is somewhat plausible[6] and that we escape Arrow's impossibility only to fall back to Kuhn's "many algorithm" problem: unless we can find a unique set of exchange rates between theoretical virtues that are rationally compelling, it looks like we have many possible theory choice rules again. But I don't think that is quite fair. Progress has been made. We are not back where we started: we now know some of the constraints on rational theory choice. They are given by the Pareto condition, the non-dictatorship condition, perhaps universal domain and at least *part* of the IIA condition (but not the part that restricts you to ordinal information).

Neither of the above listed components of the informational basis escape are immediately obvious, and despite what Okasha says in Section 7, there doesn't seem to be a compelling reason to think that scientific rationality requires that there be such objective cardinal measures of the virtues or that they be comparable. Note, however, that individual scientists do choose among theories, and such choices implicitly reveal the tradeoffs that that scientist takes to be advantageous. The next section develops this idea.

6. Theory choice at the level of the individual scientist

$R_\cap$ – which encodes the abstract, objective constraints on scientific rationality – can be incomplete. Individual scientists must ultimately make choices and thus *their* preference relations must be complete. How do they "fill in the gaps" left by $R_\cap$? Is it the case that anything goes? Or are there further restrictions on how the individual scientists fill in their personal, subjective preference relations among theories?

Ultimately, individual scientists do make choices among theories: they choose to work on one theory rather than another. That reveals that they prefer this bundle of theoret-

[6] Stegenga (2011), writing about a closely related problem, is less confident.

ical values[7] rather than another. They implicitly endorse certain tradeoffs in preferring one theory to another. Note that we are now talking in terms of the individual scientist's *subjective* assessment of the theoretical values. There needn't be some objective measure of the fruitfulness of a theory: there need only be the scientist's subjective assessment of that fruitfulness. We have moved from talking about what we can say about theory choice in the abstract to what we can say in the concrete case of a particular (but arbitrary) scientist. We have moved from talking about the *process* of aggregating the theoretical values, to talking about the *product* of that aggregation: the aggregate goodness relation. Constraints we put on this preference relation will feed back into constraints on the aggregation. We are now allowing an element of subjectivity to enter theory choice, through the ways individual scientists fill in the gaps in $R_\cap$. In doing this, we are not making theory choice a sujective, irrational choice: scientists' preferences are still subject to the rational constraints discussed earlier. It's only to the extent that those constraints fail to determine choice that subjectivity enters the picture.

The claim is that an individual scientist can use Sen's informational basis escape to successfully choose among theories. In doing so she (at least implicitly) subscribes to certain tradeoffs among the values. Can we say anything about individual scientists' preferences among theories over and above that they should be supersets of $R_\cap$? What can we say about how the scientist trades off one value for another? Luckily for us, there is already a well-developed literature on constraints on rational preference that we can appeal to. Rational choice theory is the study of what structures of preferences are rational. It is not irrational for you to strictly prefer strawberry ice cream to chocolate ice cream, but if you have that preference it is irrational for you to strictly prefer chocolate ice cream to strawberry. The theory does not pass judgement on the tastes of individuals, but it seeks to constrain the *patterns* of preferences that agents evince. If your pattern of preferences has certain structural features (transitivity, completeness...) then there is some function such that you are choosing as if you were maximising that function (Grant and Zandt 2009; Kreps 1988). If the choice environment has certain structural features then we can break down the function into component parts. A common example of this is cases where we can interpret the function you are maximising as an expected utility function and we can break it down into your probability function and your utility function. What representation theorems like this reveal about your psychology (if anything) is a matter of some debate.[8] But this much is uncontroversial: in order for there to be a maximising representation of you, there are certain properties that your preferences must satisfy. So, in order for a scientist to cardinally value the components of a bundle of theoretical values in a consistent way, she must at least satisfy the necessary conditions for a representation theorem.

For our purposes, perhaps the most relevant example is the theory of conjoint measurement (Krantz et al. 1971, Chapter 6). We can model a scientific theory as a bundle of goods. The values of the theory are the individual goods. If we represent a theory as

[7] I've moved from talking about "virtues" to talking about "values" to indicate that we have shifted focus from the abstract, objective assessment of theories to the individuals' subjective assessment. Such assessments are still constrained by the objective criteria, however.

[8] See Christensen (2001); Meacham and Weisberg (2011); Zynda (2000).

a vector of values $\alpha = (a^1, a^2, \ldots, a^n)$ then there are necessary and sufficient conditions for preference among bundles of this type to representable as maximising a function of the form $\Phi(\alpha) = \sum_i \phi^i(a^i)$. The representation is unique up to a certain class of transformations.[9] This is a special case of a more general theory for when you can be represented as maximising some function of the ϕ^is.[10] My aim here is not to champion a particular representation (and thus a particular set of constraints on rational preference among theories), but rather to point out that there is a rich tradition of finding such constraints and tying sets of constraints to numerical representations of the preferences. The point is that in assessing whether an individual agent can avoid Arrow's theorem via the informational basis escape, we can appeal to the measurement theory literature to show exactly what it would take for the agent to have (subjective) cardinal measures of the theoretical values. There are certain constraints on the scientists choice (preference) that must be fulfilled in order for such cardinal measures (the ϕ^is) to exist. It is difficult to claim that there is an objective cardinal measure of simplicity of a theory, and that is why the informational basis escape appears tricky. It is much more reasonable to assume that an individual scientist has some (at least implicit) subjective cardinal measure of the theoretical values, and we can use representation theorems to flesh out this claim.

So here we have another way that rational theory choice is constrained. The patterns of preferences among theories that individual scientists' evince can be constrained. If we want to appeal to the informational basis escape from Arrow's theorem, then we must demand that the scientists preferences satisfy certain properties. That is, for it to be possible for the scientist to assess the theoretical values in a cardinal way and to have consistent tradeoffs among them, there are certain properties that her preferences must satisfy. This bolsters what I said at the end of the last section: by taking the informational basis escape we have fallen back into Kuhn's "many-algorithm" problem, however *progress has been made.* The above discussion highlights a number of ways that scientists' choices are constrained.

Different scientists will have different preferences, so the arational tastes of actual scientists have entered theory choice, but scientists preferences are still constrained: individuals may have different preferences but they should all conform to the same *structural* constraints. The appendix lists some examples of constraints on preferences over bundles of values that might appeal. The aim here is not to argue that this or that collection of axioms are the genuine rationality constraints to put on scientists' preferences, but rather to point out that there is a literature we can appeal to in thinking about what sorts of constraints might be plausible. It might be that no such constraints deserve acceptance, and that we are therefore stuck with just $R_\cap$ as the only constraint on rational theory choice, but it might be that we can make scientific rationality more contentful by appeal to principles of preference from theoretical economics and measurement theory.

So $R_\cap$ is a constraint on rational theory choice. We might also take some principles from the theory of conjoint measurement to be structural constraints on rational preference among scientific theories. But if that's all we require then it is still open to an

[9] See the appendix for details.

[10] Certain special cases are discussed in Krantz et al. (1971, Chapter 7).

agent to put arbitrarily high and low weights on the various values and thereby to have any preference compatible with $R_\cap$. The aim of the next section is to see whether we can say more than this.

7. Trade-offs and partial commensurability

I don't know how much a watermelon costs. I also don't know how much a Japanese Yen is worth. If you were to say to me "Would you buy a watermelon for a thousand yen?" I would say "I really don't know." The two goods are incommensurable for me. If you were to say "What about five hundred yen? Two hundred?", I would still not know (but I would prefer a lower price). I would, however, be able to say "Yes, I'd buy a watermelon for one thousandth of a yen." Why? Because despite not knowing how to trade off watermelons for yen, I know that the "right" rate of exchange should lie in some range and that one thousandth of a yen is outside that range (to my advantage). So despite not being able to come up with a rate of exchange for watermelon to yen, I can recognise some trades as definitely advantageous to me, and some as definitely not advantageous to me.

I think that trade-offs between theoretical values are like this. Consider the scientist who initially prefers the simpler theory α. Imagine that some new evidence comes in that supports theory β. As evidence accrues that supports β over α, at some point, the additional accuracy of β will outweigh the scientist's initial preference for the simpler theory. Given that different individuals subscribe to different exchange rates, the scientists will move from one theory to another at different times. So scientists who complete their theory choice algorithms differently – who have different exchange rates – can reasonably disagree. As we shall see in the next section, Kuhn argues that such disagreement is a good thing. On the other hand, there will be periods where one theory is so overwhelmingly the best that however the scientists trade off one value for another, they almost all agree on what theory is best. The theory is robustly best, in a sense. This characterises the periods Kuhn calls "normal science". The end of a paradigm is characterised by the build up of anomalies. Anomalies are exactly failures of a theory to accommodate data, and are thus things that reduce the accuracy of the theory.[11] As the accuracy comes down, the scientists' differences of view about the other values cause them to abandon the old view at different times. As a new theory emerges, it begins to do better on several values and scientists are drawn towards it.

> Gradually the number of experiments, instruments, articles, and books based upon the paradigm will multiply. Still more men, convinced of the new view's fruitfulness, will adopt the new mode of practicing normal science, until at last only a few elderly hold-outs remain. And even they, we cannot say, are wrong. Though the historian can always find men – Priestley, for instance

[11] Or accuracy might be maintained at the cost of some simplicity by, for example, adding more epicycles to your Ptolemaic astronomy. Or scope might be sacrificed by claiming that your theory does not need to account for the anomalous data. In any case, an anomaly brings about some change in the bundle of values that represents the theory, and that change is detrimental.

> – who were unreasonable to resist for as long as they did, he will not find a point at which resistance becomes illogical or unscientific. (Kuhn [1962] 1992, p. 159)

Finding rationally binding exchange rates between the virtues is a losing game; but some trade-offs are still obviously worthwhile. Consider two theories α and β where $\alpha R_s \beta$ and $\beta R_a \alpha$ where R_s and R_a are simplicity and accuracy. The intersection of these relations has no relation holding between α and β. But let's imagine that the fit of β is *so much more* striking than is the simplicity of α that there is no question that β is the better theory. We can go beyond the intersection-of-the-relations relation by also considering such obviously advantageous exchange rates.

There is no rationally mandated exchange rate between the values, but some exchange rates are obviously ruled out as unreasonable, although perhaps not as contravening scientific rationality. So theory choice is rationally constrained, but not rationally determined. When particular scientists are required to make choices about which theory to pursue, they perhaps use some personal, extra-rational (but rationally constrained) principles to settle on particular exchange rates to use to make a determinate choice. Thus different scientists will choose differently.

8. Kuhn vindicated

With the help of a subtle, nuanced understanding of rationality, Okasha's use of Arrow's theorem in theory choice can be undermined. In this section, I want to point out that the view that emerges from this dialectic is more or less what Kuhn thought all along. Or rather, that my gloss on the "escape route" that Okasha discusses gives rise to an understanding of scientific rationality that is consonant with what Kuhn says. All quotes in this section are from Kuhn ([1977] 1998).

Here's Kuhn on individuals' theory choices:

> I continue to hold that the algorithms of individuals are all ultimately different by virtue of the subjective considerations with which each must complete the objective criteria before any computation can be done. (p. 109)

One could read Kuhn here as saying that each scientist must determine her own subjective (but rationally constrained) exchange rates between the values before she can work out which theory is best by her lights. We can understand the "objective criteria" that need to be completed to be those constraints encoded in $R_\cap$.

Kuhn is clear that he doesn't take the theoretical virtues he presents to *determine* theory choice, but rather, to influence it.

> Opposing maxims alter the nature of the decision to be made, highlight the essential issues it presents, and point to those remaining aspects of the decision for which each individual must take responsibility himself. (p. 110)

And later:

> I am suggesting...that the criteria of choice...function not as rules, which determine choice, but as values, which influence it. (p. 111)

Expanding on this later he says:

> Now, consider a situation in which choice by shared rules proves impossible, not because the rules are wrong but because they are, as rules, intrinsically imcomplete. Individuals must then still choose and be guided by the rules (now values) when they do so. For that purpose however, each must first flesh out the rules and each will do so in a somewhat different way even though the decisions dictated by the variously completed rules may prove unanimous. (p. 113)

A scientist fleshes out her rules by filling in the gaps in $R_\cap$.

In fact, Kuhn goes further and argues that an objective, binding rational theory choice algorithm would be a *bad thing.* It would be descriptively poor, and normatively ill-advised. On the first point:

> What the tradition sees as eliminable imperfections in its rules of choice I take to be in part responses to the essential nature of science. (p. 110)

And later:

> [T]heory choice...can be explained only in part by a theory which attributes the same properties to all the scientists who must do the choosing. Essential aspects of the process generally known as verification will be understood only by recourse to the features with respect to which men may differ while remaining scientists. (p. 113)

Scientists may differ in their subjective evaluations of the values of certain theories, but they "remain scientists" by having preferences among theories in accordance with the axioms; by having their preferences structured in the right way and agreeing with $R_\cap$.

On the second point, Kuhn says:

> [The development of new theories] *requires* a decision process which permits rational men to disagree, and such disagreement would be barred by the shared algorithm...If it were at hand, all conforming scientists would make the same decision at the same time. With standards for acceptance set too low, they would move from one attractive global viewpoint to another, never giving traditional theory an opportunity to supply equivalent attractions. With standards set higher, no one satisfying the criterion of rationality would be inclined to try out the new theory, to articulate it in ways which showed its fruitfulness or displayed its accuracy and scope. I doubt that science would survive the change. (p. 112, emphasis in original)

So it is, in fact, just as well that we were unable to develop a completely objective theory choice algorithm, since such an algorithm would be detrimental to the progress of science. There are, in these remarks, the core of the ideas of the division of cognitive labour taken up by, for example, Kitcher (1990).

Kuhn laments the use of "textbook science" in philosophy as giving a skewed picture of the development of scientific theories. Too much emphasis is given to *crucial experiments* which, while convincing, are not what actually convinced the scientists of the time.

> These [crucial] experiments are paradigms of good scientific reason for scientific choice... But... by the time they were performed no scientist still needed to be convinced of the validity of the theory their outcome is now used to demonstrate. Those decisions had long since been made on the basis of significantly more equivocal evidence. (p. 108)

Implicit in Kuhn's remarks here is the idea that a scientist who refused to be convinced by a crucial experiment would be irrational. We can imagine a crucial experiment which drastically increases the accuracy of a theory. Let's imagine that its rival theory is simpler. That the experiment is crucial means that the increase in accuracy is such that, after having performed the experiment, everyone agrees that the first theory's increase in accuracy is worth the trade-off in simplicity. To think otherwise would be to have unreasonable tradeoffs: to be like Priestly. That is, every scientists' exchange rate is such that the more accurate theory is preferred over the simpler theory.

9. Conclusion

Theory choice is a case of rational silence, not a case of incompatible constraints. This means I am siding with Kuhn's view over the view suggested by Okasha. This position requires that we allow rationality to be merely a constraint on choice, rather than the determinant of choice. There is good reason to take this position anyway. In the spirit of Okasha's goal to apply techniques from theoretical economics to problems in epistemology, I have discussed how scientists' preferences among theories might be rationally constrained using the idea that a scientific theory can be thought of as a bundle of scientific values. In fact, Kuhn argues that it would be a bad thing if scientific rationality were to determine theory choice, since this would seriously limit the valuable diversity of scientific research. I have shown how a nuanced picture of scientific rationality can allow scientists to be rational while still allowing the community of scientists to display the diversity of viewpoints that Kuhn took to be valuable.

A. Mathematical Appendix

I will briefly outline a version of Theorem 13 of Chapter 6 of Krantz et al. (1971). Let $A^1, A^2 \dots A^n$ be sets, each A^i represents the various levels of value i that a theory can take. A theory is a bundle of theoretical values, that is, a member of the Cartesian product of the A^is, which we call $\mathcal{A}$. So $\alpha = (a^1, a^2, \dots a^n) \in A^1 \times A^2 \times \cdots \times A^n = \mathcal{A}$. Scientists have preferences over theories. Let M be some subset of $N = \{1, 2, \dots, n\}$. Call $\mathcal{A}_M$ the Cartesian product of only those values listed in M. So if $M = \{1, 3\}$ then

$\mathcal{A}_M = A^1 \times A^3$. The preference over $\mathcal{A}$ is denoted $\succeq$. Let $\sim$ and $\succ$ denote the symmetric and irreflexive parts of $\succeq$ respectively. The next few paragraphs describe certain properties we might consider making constraints on rational preference among bundles of theoretical values. Together, these constraints imply a representation theorem.

Define the relation $\succeq_M$ to be a relation on $\mathcal{A}_M$ which agrees with $\succeq$ for some fixed choice of the a^is for $i \notin M$. For example, if $\mathcal{A}$ has only two components,[12] and $M = \{1\}$, then $(a^1, a^2) \succeq_M (b^1, a^2)$ if and only if $(a^1, a^2) \succeq (b^1, a^2)$ for some fixed choice of a^2. Call $\succeq$ *independent* if the relations $\succeq_M$ do not depend on the choices of the fixed components a^i for $i \notin M$. For the two component case this becomes: $(a^1, a^2) \succeq_M (b^1, a^2)$ if and only if $(a^1, b^2) \succeq_M (b^1, b^2)$ for all $b^2 \in A^2$. What independence allows is that you can determine consistent orderings on the individual values from the overall preference ordering over bundles of values. This means that a scientist can consistently separate out how good a theory is in terms of each value independently of its goodness on the other values. Recall that this was the starting point for Arrow's theorem – we started with individual values' orderings – so this condition should be reasonable. Without it, we'd be stuck with cases where a theory's simplicity affects how fruitful it is considered, for example.

Restricted solvability requires that the following conditional be satisfied. If there exist $a^i, b^i \in A^i$ for all $i \neq j$ and $\overline{b^j}, \underline{b^j} \in A^j$ such that:

$$(b^1, \ldots, \overline{b^j}, \ldots, b^n) \succeq (a^1, \ldots, a^j, \ldots a^n) \succeq (b^1, \ldots, \underline{b^j}, \ldots, b^n)$$

then there exists $b^j \in A^j$ such that

$$(b^1, \ldots, b^j, \ldots, b^n) \sim (a^1, \ldots, a^j, \ldots a^n)$$

This means that if there are theories $\overline{\beta}$ and $\underline{\beta}$ that differ only in one component and where one is better than and one worse than α, then there is a theory β that differs from $\overline{\beta}$ and $\underline{\beta}$ only in that same component which is just as good as α. What this requires is that the possible levels of the values are suitably rich. Restricted solvability is really about the space of possible theories that we require a scientist to have preferences over. The idea is that if there are these theories that are better and worse than α just in virtue of their simplicity, say, then there is a hypothetical theory that is exactly as good as α. There should not be a "simplicity gap" in "theory space".

Let X be some (finite or infinite) set of consecutive integers. A standard sequence for value 1 is a set $S = \{a^1_i | i \in X\}$ with the following property. For $p, q \in \mathcal{A}_M$ where $M = N \setminus \{1\}$ and it is not the case that $p \sim_M q$, we have $(a^1_i, p) \sim (a^1_{i+1}, q)$ for all $i, i+1 \in X$. There's a little abuse of notation there: (a^1_i, p) is the element of $\mathcal{A}$ that has a^1_i as its first component and the respective components of p in each other place. A standard sequence is basically a "yardstick" for measurement. For example, a standard sequence for length is "A 1cm stick, a 2cm stick, a 3cm stick…". The idea is that p and q identify some distance between consecutive members of the standard sequence (for the length case, 1cm) and the a^1_i, a^1_{i+1} trade off that difference between p and q. A standard

[12] The theorem I am discussing only works if there are at least three components: the two component case works slightly differently. The discussions of the two component case here are for illustrative purposes only. See Krantz et al. (1971, Chapter 6) for details.

sequence S is bounded if and only if there exist $b^1, c^1 \in A^1$ such that $b^1 \succ_M a_i^1 \succ_M c^1$ for every $a_i^1 \in S$ where $M = N\backslash\{1\}$. Standard sequences for other components are defined in the same way. The relation $\succeq$ satisfies the *Archimedean axiom* if every strictly bounded standard sequence is finite. This axiom requires that there are no standard sequences with infinitessimally small steps. It also rules out cases of lexicographic orderings.

A component A^i is *essential* if there are $a_1^i, a_2^i \in A^i$ and $p \in \mathcal{A}_M$ where $M = N \setminus \{i\}$ such that it is not the case that $(a_1^i, p) \sim (a_2^i, p)$. We are again using the same abuse of notation mentioned in the last paragraph. A component is essential when its value can tip the balance in favour of one theory or another. Kuhn listed five values, Longino (1996) gives an alternative list of values: it's reasonable to think that at least three will be essential in this sense. One could have the same discussion here that Okasha had about the Non-dictatorship condition, but since the discussion would be the same, I omit it.

$\succeq$ has an *Additive Conjoint Representation* if there exist functions $\phi^i \colon A^i \to \mathbb{R}$ such that the following biconditional holds. $(a^1, \ldots, a^n) \succeq (b^1, \ldots, b^n)$ if and only if $\sum_i \phi^i(a^i) \geq \sum_i \phi^i(b^i)$. If the ψ^is are also an additive conjoint representation of $\succeq$ then there exist real numbers x, y_i such that $\psi^i = x\phi^i + y_i$. Note that it is the same x for each component. This means that the components are *unit comparable*. This is a kind of comparability that is less than full comparability (Okasha 2011, p. 100).

The representation theorem is then the following. If $\succeq$ is a complete, transitive and reflexive relation on $\mathcal{A}$ which satisfies Independence, Restricted Solvability, Archimedean Axiom and at least 3 components are essential then there is an additive conjoint representation of $\succeq$. The proof is on pp. 307–9 of Krantz et al. (1971), although details necessary to understand the proof are spread out through several earlier chapters.

References

Christensen, David (2001). "Preference-based arguments for probabilism". In: *Philosophy of Science* 68, pp. 356–376.

Gaertner, Wulf (2009). *A Primer in Social Choice Theory.* Oxford University Press.

Gaifman, Haim (1983). "Paradoxes of Infinity and Self-Applications, I". In: *Erkenntnis* 20, pp. 131–155.

Grant, Simon and Timothy van Zandt (2009). "Expected utility theory". In: *The handbook of rational and social choice.* Ed. by Paul Anand, Prasanta K. Pattanaik, and Clemens Puppe. Oxford University Press, pp. 21–68.

Kitcher, Philip (1990). "The Division of Cognitive Labor". In: *Journal of Philosophy* 87, pp. 5–22.

Krantz, D et al. (1971). *Foundations of Measurement Volume I: Additive and Polynomial Representations.* Dover Publications.

Kreps, David M. (1988). *Notes on the Theory of Choice.* Westview Press.

Kuhn, Thomas ([1962] 1992). *The Structure of Scientific Revolutions, 3rd edition.* University of Chicago Press.

Kuhn, Thomas ([1977] 1998). “Objectivity, value judgment and theory choice”. In: *Philosophy of Science, The Central Issues*. Ed. by Martin Curd and J.A. Cover. W.W. Norton and Company, pp. 102–118.

Longino, Helen (1996). “Cognitive and Non-cognitive Values in Science: Rethinking the Dichotomy”. In: *Feminism, Science and the Philosophy of Science*. Ed. by L.H. Nelson and J. Nelson. Kluwer Academic Publishers, pp. 39–58.

Meacham, Christopher and Jonathan Weisberg (2011). “Representation Theorems and the Foundations of Decision Theory”. In: *Australasian Journal of Philosophy* 89, pp. 641–663.

Okasha, Samir (2011). “Theory choice and social choice: Kuhn versus Arrow”. In: *Mind* 477, pp. 83–115.

Stegenga, Jacob (2011). “An impossibility theorem for amalgamating evidence”. In: *Synthese*.

Zynda, Lyle (2000). “Representation Theorems and Realism about Degrees of Belief”. In: *Philosophy of Science* 67, pp. 45–69.