The Rationality Principle Idealized

Boaz Miller

Published in *Social Epistemology* 26(1): 3-30, 2012.

Abstract: According to Popper's rationality principle, agents act in the most adequate way according to the objective situation. I propose a new interpretation of the rationality principle as consisting of an idealization and two abstractions. Based on this new interpretation, I critically discuss the privileged status that Popper ascribes to it as an integral part of all social scientific models. I argue that as an idealization, the rationality principle may play an important role in the social sciences, but it also has inherent limitations that inhibit it from having the privileged status that Popper ascribes to it in all cases.

Keywords: abstractions, explanation, idealizations, models, Popper, rationality, social sciences.

Introduction

Sir Karl Popper's 'rationality principle' (RP) constitutes one of the corner stones of his philosophy of the social sciences, along with situational analysis and methodological individualism. According to RP, agents act in the most adequate way according to the objective situation. Popper maintains that RP has minimal empirical content, yet he elevates RP to the status of the animating law of all models in the social sciences. RP has been subject to different interpretations, and some philosophers have found it difficult to reconcile RP and the privileged status that Popper ascribes to it with his commitment to critical rationalism.

Popper presents RP in the context of his discussion of the role of models in science in general and the social sciences in particular. Philosophical discussions of RP have largely overlooked the importance of this context to understanding RP. Drawing inter alia on recent literature about models in science, I propose a new interpretation of RP as consisting of an idealization and two abstractions. My interpretation is an alternative to the common statistical and methodological interpretations of RP. Based on my new interpretation, I critically discuss the privileged status that Popper ascribes to RP. I argue that as an idealization, RP may indeed

play an important and often indispensable role in the social sciences. However, my analysis also reveals that RP has inherent limitations that inhibit it from having the privileged status that Popper ascribes to it in all cases.

This paper consists of six sections. In Section 1, I analyze Popper's view of models in science. This section sets the stage for the rest of the discussion in the paper. In Section 2, I reject a common interpretation of RP as a statistical law. In Section 3, I argue that for Popper, idealizations and abstractions, as they are employed in models, are true representations of aspects of the world. With respect to irrelevant factors that abstractions and idealizations omit in their representations of the world, I distinguish between three types of irrelevance: domain irrelevance, causal irrelevance, and concomitant irrelevance. In section 4, I present my interpretation of RP as consisting of an idealization and two abstractions. In section 5, I discuss the difficulty of RP with explaining cases in which people behave in a seemingly irrational way. I illustrate this difficulty with Kahneman and Tversky's prospect theory. In section 6, I argue that my interpretation of RP as an idealization resolves this difficulty and preserves the privileged status Popper ascribes to it in the social sciences, except in cases in which we are interested in explaining the aspects of the models from which RP abstracts away.

1 Popper on Models in Science

Popper discusses RP in the context of his discussion of models in science. He begins with a discussion of the role of models in the natural sciences, from which he develops his account of their role and the role of RP in the social sciences.[1] In this section I will critically discuss Popper's account of models in science. My aim is not to evaluate the correctness of his account,

[1] The discussion of Popper's account of models draws on two main sources. The first is chapter 29 of The poverty of historicism (Popper [1957] 1961, 130-143), and the second is his paper 'Models, instruments, and truth: The status of the rationality principle in the social sciences' (Popper 1994). While published in its final form only in 1994, it is based on a lecture delivered in 1963, to which more sections were added later. An extract of it was published in French (Popper 1967) and in English (Popper 1985). All of my references to this paper will be to its latest version, which contains important additions.

but rather to highlight key aspects of it, which will be relevant for the discussion in the subsequent sections. My argument in this section is that Popper regards models, which he takes to be abstract or concrete objects, as a genuine subset of theories, rather than merely useful heuristic devices.

Two distinct functions of models in science may be identified in Popper's account. First, models mediate observation. We observe the world through the lens of our models. Second, models help us explain certain things in the world:

> And in the social sciences it is even more obvious that we cannot see and observe our objects before we have thought about them. For most of the objects of social science, if not all of them, are abstract objects; they are theoretical constructions. (Even 'the war' or 'the army' are abstract concepts, strange as this may sound to some. What is concrete is the many who are killed; or men and women in uniform, etc.) These objects, these theoretical constructions used to interpret our experience, are the result of constructing certain models (especially of institutions), in order to explain certain experiences – a familiar theoretical method in the natural sciences (where we construct our models of atoms, molecules, solids, liquids, etc.) (Popper [1957] 1961, 135; emphasis in the original).[2]

How do models help us gain access to the world? They help us by representing the objects in the world that we want to study. Popper argues that models are especially useful in cases in which we want to explain or predict kinds of events, such as the repeating occurrence of lunar eclipses. If we want to account for the repeating occurrence of lunar eclipses, it would be useful to build a model of the solar system (which can even be a concrete physical model), and 'animate' the model or set it in motion using the laws of our theory, in this case Newton's laws (Popper 1994, 161-2). He writes:

[2] An interesting question that arises from this quote and exceeds the scope of this paper is Popper's view on the reality of the social institutions that are represented by models. For a strong ontological reading of Popper's position, which criticizes him for denying the existence of social institutions, see Winch (1958) at 126-128. For an interpretation of his position as not denying the existence of institutions, but rather as denying only social wholes that are irreducible to individuals and their actions, see Agassi (1960) and Jarvie (2001) at 126-129.

> Models represent typical initial conditions rather then universal laws. And they therefore need to be supplemented by 'animating' universal laws of interaction – by theories which are not models in the sense here indicated (Popper 1994, 165).

The last sentence draws a distinction between models and theories. Models are concrete or abstract objects that represent typical relations, while theories are sets of universal laws that set models in motion. He adds:

> Models […] may be called 'theories', or be said to incorporate theories, since they are attempts to solve problems of explanation. But the opposite is far from true. Not all theories are models (Popper 1994, 165; emphasis in the original).

The sense in which models 'incorporate' theories is clear from the above analysis, namely models such as the model of the solar system incorporate theories as their universal animating laws. But in what sense are models a subset of theories, as the last sentence implies?[3] A comparison between the models of the solar system and the atom, both of which Popper mentions, may shed light on this matter. There are two relevant differences between the model of the solar system and that of the atom. First, in the model of the solar system, the arrangement of the planets is independent of the Newtonian theory, understood as a set of laws. The distances of the planets from the sun and from one another, their number, their masses, etc. are all contingent. The same Newtonian laws could operate in a universe in which the planets were arranged differently. By contrast, in the model of the atom, Bohr's theory dictates the distances of the electron from the nucleus. The theory states that electrons can only occupy certain places that correspond to certain energy levels. The number of electrons in each energy level is also dictated by the theory. In other words, in the model of the atom, the

[3] While it is clear, in my opinion, from this passage that for Popper some but not all models are theories, Notturno interprets Popper in this passage as claiming that models are not genuine theories: 'Model may be called theories, but real theories represent abstract universal laws, whereas models represent typical (and not necessarily actual) initial conditions (Notturno 1998, 407; emphasis in the original). In my view, not only does my interpretation better accord with the literal interpretation of this passage, it is also preferable for better according with Popper's notion of approximate truth and his view about the unity of method, as I will argue toward the end of the end of this section.

theory does not only supply the 'animating laws', but also imposes constraints on the way the atom is represented.

Second, the model of the solar system is a model of our concrete solar system. By contrast, the model of the atom is not a model of any specific atom, such as the hydrogen atom. Rather, it is a general model, which can instantiate more specific models. It is general in scope, and applicable to a large number of cases. In this sense, it resembles more a theory, as it does not represent a concrete object in the world, but rather generalizes over objects of a certain kind. In the case of the model of the atom, as opposed to the model of the solar system, the distinction between what counts as pure representation and what counts as the animating laws is blurry. Because of that, the model of the atom may be legitimately referred to as the theory of the atom.

This analysis shows why Popper may regard models as a subset of theories. However, there seems still to be a problem of a difference in kind between models and theories. For Popper, theories are linguistic entities, while models are abstract or concrete objects. As far as I have checked, Popper does not address this difficulty. However, this difficulty is addressed by Kenneth F. Schaffner (1993). In my view, Schaffner's view is an implicit assumption in Popper's account of models. Therefore, I will briefly discuss it in the remainder of this section.

Schaffner suggests an account of explanation in the biological and medical sciences, which is a variant of the deductive-nomological (D-N) account of explanation (Hempel & Oppenheim 1948). However, he notes a difficulty with his suggestion, which is that most biomedical theories do not conform to the D-N structure. Rather, Schaffner characterizes biomedical theories as sets of models, such as protein synthesis models, and biochemical models. Schaffner notes two main differences between these sets of models and theories in the D-N form. First, a model is usually not represented as a set of generalizations, and even when it

is, such generalizations are usually very domain-specific. Since they include descriptions of specific characteristics of a particular target system, they already include what is considered as 'initial conditions' on the D-N account. Second, generalizations in biomedicine are usually only analogous to the actual mechanisms in the different organisms. This is because they generalize over mechanisms in different organisms or members of the same organism. Such mechanisms have small but potentially significant differences, which have evolved through natural selection (Schaffner 1993, 285).

In spite of these differences between theories understood as sets of models and theories understood as linguistic entities in the D-N form, Schaffner still makes the following claim, which, in my view, is an implicit assumption in Popper's account of models:

> All of this said, however, there is no reason that the logic of the relations between the explanans and the explanandum cannot be cast in deductive form. Typically this is not done because it requires more formalization than it is worth, but some very complex engineering circuits […for example] can effectively be represented in the first order predicate calculus and useful deductions made from the premises… (Schaffner 1993, 287-28; emphasis in the original).

There is no need to discuss the correctness of Schaffner's claim here.[4] It suffices to note that Schaffner provides the missing link in Popper's account of models, namely the assumption that the transition between a formulation of a theory in the form of a set of abstract or concrete objects to its formulation in the D-N form is possible and unproblematic.

This implicit assumption is critical to Popper's account in yet another way. Popper rejects an instrumentalist view of models in the social and natural sciences, and defends a realist one. Popper claims that successful models are approximately true. The notion of truth approximation on which Popper relies in this context is his theory of verisimilitude (Popper

[4] Schaffner's example seems question-begging, as the logical circuits he describes are by definition a graphical representation of first-order logical propositions.

1994, 173-175).[5] As Popper's theory of verisimilitude is defined in terms of theories formulated in the D-N form, it is vital for him that models can be expressed in this form.

Last, without adding this implicit assumption to Popper's account, there would be a tension in his view about the unity of method between the natural and social sciences. As I will explain in the next sections, Popper maintains that the unlike the natural sciences, the 'animating law' of models in the social sciences has only minimal content, whereas the model does the rest of the explanatory work. If we restrict theories to abstract universal laws, i.e. maintain that models are not genuinely theories, it may follow that the social sciences are not genuinely science. This conclusion seems inconsistent with Popper's view about the unity of science, and should therefore be rejected.

From his general discussion of models in science, Popper turns to discuss models in the social sciences. Popper stresses that models represent reality in a simplified form. Popper also maintains that the law that animates models in the social sciences is RP. In the next section, I argue against a common interpretation according to which RP is a statistical law, and that this is the reason why it helps scientists construct approximately true models. This will enable me to propose my alternative interpretation.

2 Is RP a Statistical Law?

Popper equates a model in the social sciences with a description of a typical social situation. A model in the social sciences is a "situational analysis of a social situation", or in short "situational logic". Popper gives an example of a man named Richard, who tries to cross the road. In order to explain Richard's pattern of movement across the road, we need to describe the situational elements that constrain his behaviour. First, there are physical bodies such as

[5] Popper writes: "...with the help of Tarski's concept of truth, [...] I think I have been able to give a purely logical definition of the relation 'a is a better approximation to the truth than b', or 'a is more similar to the truth than b'" (Popper 1994, 175). In a footnote, Popper refers the reader to David Miller's criticisms of his truth theory (Popper 1994, 183 fn 17).

cars. Then, there are social institutions, which may take the concrete form of persons, e.g. a policeperson, or physical objects, e.g. traffic signs. They can also take an abstract form, e.g. the rules of the road. To complete the description of the social situation, we must ascribe aims to Richard, such as safely crossing the road. Then, we must ascribe to him knowledge of the situation. Finally, we need to assume that Richard is acting in accordance with the situation. Popper calls the latter assumption "the rationality principle".

For Popper, RP is the 'animating law' of the model in the same way that Newton's laws are the animating laws of the model of the solar system (Popper 1994, 166-168). According to Popper, the actual content of RP is minimal – the mere assumption that 'the actors act within the terms of the model, or that they "work out" what was implicit in the situation' (Popper 1994, 169; emphasis in the original). Nevertheless, it is also a powerful assumption, as it the one that sets the model in motion.

Popper stresses the objectivity of situational analysis. According to Popper, the actions and aims that we ascribe to the actors are part of the objective social situation. He sharply distinguishes RP from psychologism. Psychologism refers to two interconnected notions though Popper himself does clearly distinguish between them. The first, which we may call "psychological nomologism", is the view that explanations of social situations appeal, eventually, to psychological laws. Namely, the "animating laws" of models in the social sciences are psychological laws. The second, which we may call "mentalism", is the view that social explanations require descriptions of individuals' mental states, such as their particular motivations and beliefs.[6] Popper rejects both these forms of psychologism, as he regards psychological laws to be unfalsifiable claims about what is going on in the actors' minds. Rather, he suggests that all that is needed for explanation is the adequacy of the actors' actions

[6] I thank an anonymous referee for this distinction.

with respect to the aims that are part of the social situation. Psychological assumptions about the actors are not needed (Popper 1994, 167-170; [1957] 1961, 141-142).

What is the status of RP? Popper says:

> The adoption of the rationality principle can therefore be regarded as a by-product of a methodological postulate. It does not play the role of an empirical explanatory theory, of a testable hypothesis. For in this field, the empirical explanatory theories or hypotheses are, rather, our various models, our various situational analyses (Popper 1994, 169).

These words may be interpreted in two different manners, both of which are problematic. According to one interpretation, RP is a methodological principle, which is not empirically tested but is rather valid a priori and is suggested by Popper as part of a successful methodology. This interpretation, however, seems to be at odds with Popper's general commitment to critical rationalism and falsificationism. On a second interpretation, RP is an empirical conjecture. However, in this case, it seems that it should be tested along the rest of the theory of which it is part, and rejected if found to be false. This is at odds with Popper's claim that RP has a privileged status as the 'animating law' of all models in the social sciences.

Popper endorses the second interpretation of his position, but rejects its alleged consequences. According to Popper, RP is an empirical claim, which is not only falsifiable, but, strictly speaking, false. Though false, it is 'as a rule sufficiently near to the truth' (Popper 1994, 178). Employing RP, more than any other method, helps scientists construct approximately true models. Therefore, adopting RP is a good methodological decision one makes when constructing a model in the social sciences (Popper 1994, 177-178).

In the rest of this paper, I will unpack these claims. I will suggest a new interpretation of what Popper means when he claims that RP is 'sufficiently near to the truth' and in what way it helps scientists construct approximately true models. My interpretation will allow me to critically examine in sections 5 and 6 the privileged status that Popper ascribes to RP.

Why does Popper claim that RP is false? He does so because of the trivial fact that people do not always act in accordance with the situation. Popper mentions two types of cases in which people seemingly violate RP. First, people may act in an irrational manner in accordance with some psychological mechanism (the development of which, though, may be rationally explained). Popper gives the example of a driver who desperately tries to park his car when there is clearly not enough parking space (Popper 1994, 172). Second, people may seem to act irrationally, but in fact act rationally in accordance with the subjective way they see the situation, which is different from the way the situation is seen objectively (Popper 1994, 178). Referring back to Popper's example of Richard trying to cross the road, we can imagine a case in which Richard, who is in a hurry, fails to notice the traffic light and crosses the road in a red light. (I will analyze these two types of cases more closely in section 5.)

In what way, then, is RP sufficiently near to the truth, and how does it help scientists construct approximately true models? A common view, which I will challenge, interprets Popper as stating a statistical law. The main defender of this view is Maurice Lagueux, who argues that 'all that Popper means when he says that the principle is false is that it is occasionally contradicted' (Lagueux 1993, 470; emphasis in the original). Referring to the example of the frustrated driver, Lagueux remarks: "Clearly, if Popper maintains that RP is still a 'good approximation,' it is because he considers that such cases are not very representative" (Lagueux 1993, 470). Lagueux interprets Popper as stating a statistical law, which is true in most but not necessarily all cases, and is therefore, strictly speaking, empirically false. However, since the principle is true in the majority of cases, models that employ it also describe the majority of cases, and are therefore approximately true.[7]

[7] Other scholars seem to share this view. For example, Noretta Koertge states: "Popper almost surely believes that RP is at best a statistical law" (Koertge 1979, 87), and Herbert Keuth takes RP to include the assumption that "people mainly act rationally" (Keuth 2005, 200).

In my view, the statistical interpretation of RP is incorrect. I argue that Popper's rationality principle is not a claim about what happens in the majority of cases. Consequently, this is neither the reason why it helps scientists construct approximately true models, nor the reason why Popper ascribes to RP a privileged status in model constructing in the social sciences.

I find the statistical interpretation implausible for two reasons. First, if all that Popper meant were that people only occasionally did not act in accordance with the situation, he could simply say so. When referring to the question of whether people act rationally or not, Popper does not talk about what happens in the majority of cases. For example, Popper says that due to vast differences between people, 'some people will act appropriately and others not' (Popper 1994, 172), without specifying the relative size of each group of people. Popper also says that people act 'more or less rationally' (Popper [1957] 1961, 140-141). The latter assertion is compatible, for example, with the view that all people act partly rationally in all cases. In fact, Popper himself seems to reject the statistical interpretation: 'if you look upon this so-called rationality principle, […] then you will find that it has little or nothing to do with the empirical or psychological assertion that man always, or in the main, or in most cases, act rationally' (Popper 1994, 169).[8]

Furthermore, claiming that what he means by RP is merely to state a statistical law would save Popper from a lot of trouble in his argument. If Popper argued that people only occasionally did not act in accordance with the situation, he would arguably be making a true empirical claim. Thus, he could defend much more easily the privileged status of RP in a manner which is consistent with his falsificationism. Thus, he could avoid the trouble of having

[8] Lagueux interprets this passage as referring to a different notion of rationality from the one that is referred to by RP. (Lagueux 1993, 472-473). Though I find no evidence in the text to support this claim, my argument does not hinge on the question of the correct interpretation of this passage.

to claim that RP is a false empirical claim, the adoption of which has nevertheless methodological merits. The fact that Popper defends a more complex position suggests, in my view, that he does not merely wish to point out a statistical regularity.

Lagueux finds support for the statistical interpretation of RP in Popper's claim that models in the social sciences explain typical social situations, but do not predict specific events (Lagueux 1993, 475). However, Popper does not ascribe this inability to the fact that statistical laws have exceptions. Rather, he sees this inability as a common state of affairs in science in general. Popper says that this inability to predict is a result of the complexity of phenomena that cannot be fully represented in simplified models, and is common to both the physical and social sciences. According to Popper, our inability to predict the behaviour of a particular individual agent results from similar reasons that are responsible for our inability to predict a particular thunderstorm or fire (Popper [1957] 1961, 139).

3 On the Truth in Abstractions and Idealizations and Three Types of Irrelevance

If not a statistical law, then, what does RP state, and in what sense is it sufficiently near to the truth? Let us begin answering this question by looking at Popper's rejection of psychologism. By rejecting mentalism, Popper claims that actors' psychological motivations are irrelevant to the method of situational logic. According to Popper, we can construct true or approximately true models of human behaviour that do not take into consideration the actors' proclaimed psychological motivations, even when such proclaimed motivations are at odds with the objective aims that the model ascribes to the actors. Popper considers the case of the economic theory of profit maximization, which states that the businessperson maximizes his or her profits by a policy of marginal cost pricing. Evidence from questionnaires that shows that businesspersons report different motivations for their actions has led philosophers to adopt an instrumentalist view of the theory as a merely useful tool for prediction, which does not

necessarily describe the real behaviour of businessmen. By contrast, Popper adopts a realist interpretation of this theory. He writes:

> Against all this, I suggest that the method of situational logic is not connected with what the agent's actual thoughts were when performing the actions (compare the case of Richard's crossing the street). In consequence, evidence from questionnaires about psychological motivation is not necessarily relevant to the testing of a theory about situational logic (Popper 1994, 182).

The claim here is similar to Popper's claim that all that is needed to explain Richard's crossing the street is the social institutions, his objective aims and knowledge of them, and the principle of adequacy. Richard's knowledge of Verdi's operas or Sanskrit texts is irrelevant to the situation, even if while crossing the street he was humming a passage from Verdi or thought about the Atharva-veda (Popper 1994, 168).

What this shows is that for Popper, RP serves as a criterion for excluding parts of the story which Popper deems irrelevant to the explanation. In proposing RP, Popper makes a qualitative distinction between different elements of the social situation based on their relevance to the explanatory model. I therefore propose that RP be understood as near to the truth in the sense of telling only parts of the story, and not in the sense of being occasionally false, as the statistical interpretation suggests.

An analogy Popper draws between models in the physical and social sciences supports my interpretation. Popper states that models oversimplify the facts. They are therefore false but also approximately true. A model constructed in the social sciences using RP is not different from a model in the physical science, such as the model of the solar system, about which Popper says:

> Take the Newtonian model of the solar system. Even if we assume that Newton's laws of motion are true, the model would not be true. Though it contains a number of planets – in the form, incidentally, of mass-points, which they are not – it contains neither the meteorites nor the cosmic dust. It contains neither the pressure of the light of the sun nor that of cosmic radiation. It does not even contain the magnetic properties of the planets, or the electrical fields that result in their neighbourhood from the movements

> of these magnets. And – perhaps most important – it does not represent the action of distant masses upon the bodies of the solar system. It is, like all models, a vast over-simplification. (Popper 1994, 172).

Note that models in physics are not false because they describe what happens only most of the times. Rather, as this quote shows, they are, strictly speaking, false because they systematically omit certain elements of physical reality (meteorites, magnetic fields, etc.) and systematically misrepresent other elements of it (planets as point-masses). I argue that RP is an oversimplification of the same type that Popper identifies in the physical sciences. RP helps construct approximately true models in the same way that systematic omissions and misrepresentations help construct models in the physical sciences.

This analogy militates against another interpretation of RP. Mark A. Notturno argues that RP is 'a methodological principle that places restrictions on what will and will not count as a rational explanation'. He claims that RP is a precondition for explanation in the social sciences in the same way that the methodological postulation that there are laws of nature (even if false) is a precondition for rational explanation is the natural sciences (Notturno 1998, 405). However, even if on Notturno's interpretation, RP avoids the charge of being an a priori metaphysical assumption,[9] this interpretation seems wrong. As Popper states, RP is the animating law of a situational model in the same way Newton's laws are the animating laws of the model of the solar system. Put differently, RP is a law, not a precondition for explaining with a law. The crux is that in the same way that Newton's laws can be used to describe the situation truthfully only when certain elements of the situation are oversimplified – omitted or misrepresented – so can RP.

In what ways do models oversimplify – systematically omit and misrepresent? A common distinction exists between abstraction and idealization. Martin R. Jones distinguishes

[9] See Miller (2006) at 129-130 for a defence against this charge.

idealizations, which are representations of the world that systematically misrepresent things as simpler than they actually are, from abstractions, which are representations that omit certain features of their target object. Under Jones' account, idealizations and abstractions are mutually exclusive (Jones 2005, 174-175).

Jones distinguishes between two forms of omission. One form of omission is 'omitting by representing as absent', which is involved in idealization. In this case, when we omit a feature of the target system, for example a plane without friction, we also misrepresent it. This is because there exist no frictionless planes in the world. The second form of omission is 'omitting by not mentioning', which is involved in abstraction. In this case, we omit a feature of the target system by keeping silent about it (Jones 2005, 189 fn 35).

Are abstractions and idealizations true or false? On the one hand, as we have seen, Popper regards them as false empirical claims, which enable us to construct approximately true models (under his theory of verisimilitude). On the other hand, he regards them as oversimplifications of reality, which means that they must correctly correspond, in some manner, to reality. This tension suggests that under Popper's account, abstractions and idealizations can be also considered as true or partly true in some sense.[10]

As Yemima Ben-Menahem argues, the 'either-or' categorical distinction between truth and falsehood is not helpful for dealing with idealizations. Rather, it is better to think of truth and falsehood as two extremes on a continuum, and as idealizations as more true than false. In response to Nancy Cartwright (1983), who argues that although the fundamental laws of physics lie, they explain, Ben-Menahem argues that by not distinguishing between degrees of falsehood, Cartwright offers no basis for granting an explanatory role to abstracted physical

[10] I do not mean that they are close to the truth according to a formal definition of approximate truth, such as Popper's theory of verisimilitude, but rather in a qualitative way that distinguishes them from merely false claims.

laws and denying it from mere falsities (Ben-Menahem 1988, 167-168). Ben-Menahem says: 'what should be emphasized about an idealization is not that it is, strictly speaking, false, but that it bears very special relations to what is true' (Ben-Menahem 1988, 169). I suggest that this, or something very similar, is what Popper means by saying that RP is sufficiently near to the truth, and may explain why Popper believes it helps scientists construct approximately true models.

What reasons are there for thinking that abstractions and idealizations are more true than false? Ernest Nagel gives some prima facie compelling reasons. Within the assumptions of a scientific theory, Nagel identifies three subgroups of 'unrealistic assumptions'.[11] Under the terminology I have adopted in this paper, Nagel's first subgroup corresponds to abstractions. The second subgroup consists of statements that are considered false or at least highly improbable on the available empirical evidence. The third subgroup corresponds to idealizations (Nagel 1963, 214-215).

Nagel argues that abstractions and idealizations are true (provided that the theory is true), whereas statements that belong to the second subgroup are false. As for abstractions, Nagel argues that 'no finitely long statement can possibly represent the totality of traits embodied in every concretely existing thing' (Nagel 1963, 214). In other words, since there are potentially infinitely many things to say about something, every statement will be partial with respect to the full description of the thing to which it refers. Abstractions are therefore not false. Nagel's observation is consistent with Jones' claim that an abstraction with respect to

[11] Within a scientific theory, Nagel distinguishes between three sets of statements. The first is a set of assumptions. The second is the set statements that are logically deducible from the first set. The third set is a set of rules of correspondence that relate the observable phenomena to theoretical terms that appear in the assumptions. Theoretical terms are terms that refer either to idealized notions that do not actually exist in reality, such as 'vacuum', or to unobservable entities, such as 'genes' (Nagel 1963, 212-213).

some property does not get us nearer to or further away from the truth, as it is silent with respect to whether the target system has this property or lacks it (Jones 2005, 188).

However, one may argue that a statement that fails to mention an important or a relevant part of the target system is a false statement. As we know from the legal context, a false story can be constructed from selectively choosing certain true facts, while neglecting others. This is why a witness in the legal context is required to say 'the truth, the whole truth, and nothing but the truth'. If a thief neglects to mention his theft while giving a detailed account of all the rest of the things that he did on a certain day, he is lying.

In order to address this objection, we first need to have a clear understanding of what a relevant fact is, the omission of which will cause a statement to be false. I distinguish between three types irrelevance with respect to descriptions of things in the world that may occur in a representation of a target system.

Domain Irrelevance: Objects or properties thereof that are not considered part of the target system.

For example, statements about the prime minister of Italy are generally not considered relevant to the model of the simple pendulum, and the fact that statements in the model of the simple pendulum are silent with respect to the prime minister of Italy does not affect their being more or less true with respect to pendulums. In fact, in order to avoid trivializing the notion of abstraction, Anjan Chakravartty talks about abstraction only with respect to features that are relevant to the target system in the first place (Chakravartty 2007, 190).[12]

[12] The claim that a statement's failing to mention domain-irrelevant things does not affect its truth value seems almost trivial. However, there are aspects of reality that are not considered relevant to a theory or a model, but are only later discovered to be in fact relevant. For example, Paul Feyerabend shows that the in the beginning of the twentieth century, the phenomenon of Brownian motion was considered domain-irrelevant to thermodynamics. At that time, there were two rival thermodynamical theories, the kinetic theory and the phenomenological theory. Much to their surprise, scientists discovered that Brownian motion was implied by kinetic thermodynamics. Only after the phenomenon of Brownian motion was reinterpreted in terms of the dynamic theory (as a perpetual motion machine) was its rival theory of phenomenological thermodynamics

Causal Irrelevance: Objects or properties thereof that are considered part of the target system, but do not causally affect the aspects of the behaviour of the target system that we want to describe.

Take, for example, the model of the simple pendulum. When we use the model of the simple pendulum, we are interested in describing the motion of pendulums. There are properties of pendulums, such as the mass of the weight and the length of the cord that causally affect the motion of the pendulum. The more we specify such factors, the more accurate a description we get. The less we specify such factors, the more abstract a model we have. By contrast, there are properties that do not causally affect the motion of the pendulum, and therefore we can stay silent about them. We can abstract away from them. For example, ceteris paribus the color of the pendulum as such is not causally relevant to describing its movement.

Cartwright argues that we can only abstract away from irrelevant factors, such as the color of the pendulum. However, if a factor is relevant to the behaviour of the system, such as friction, we must say something about its presence or absence (Cartwright 1989, 187).[13] If we remember Jones' distinction between omitting by representing as absent, which is part of an idealization, and omitting by not mentioning, which is part of an abstraction, then Cartwright's claim is that we cannot omit-by-not-mentioning factors that causally affect the aspects of the behaviour of the target system that we want to describe.

Jones raises two objections to Cartwright's claim. First, as in the case of domain-irrelevance, we may find out that factors that we deem causally irrelevant to the behaviour of

---

shown to imply its negation, whereas up to that point it had been considered irrelevant to it in the sense of neither implying Brownian motion nor its negation. The scientific community perceived this as a triumph of the dynamic theory (Feyerabend 1981, 71-72).

[13] My distinction between domain-irrelevance and causal irrelevance helps resolve an alleged tension between Chakravartty, who suggests defining abstraction only with relation to relevant factors, and Cartwright, who argues that we can only abstract away from irrelevant ones. It seems Chakravartty's notion of irrelevance refers to domain-irrelevance, while Cartwright's refers to causal irrelevance.

the system are in fact relevant (Jones 2005, 189).[14] Second, and more significantly, we can abstract away from some causally relevant factors in the system because our model already contains an idealization that 'screens them off'. In other words, Jones describes a situation in which an idealization that already exists in the model allows us to be silent with respect to a relevant causal factor that affects the behaviour of the system. For example, in the ideal gas model, we idealize the gas molecules to be perfectly elastic spheres which exert no attractive forces on one another. This idealization allows us in turn to stay silent about the internal structure of the molecules – to abstract away from it. This is although the internal structure of the gas molecules casually affects the intermolecular forces between them, which causally affect the gas pressure and volume (Jones 2005, 189-190). Jones' argument therefore helps us identify a third type of irrelevance:

Concomitant Irrelevance: Objects or properties thereof that are considered part of the target system, causally affect the aspects of the behaviour of the target system that we want to describe, but are screened off by an existing idealization in the model.

Having drawn this distinction, we can now see that abstractions that abstract away from either domain-irrelevant or causally irrelevant factors may be true. But what about abstractions that abstract away from concomitantly irrelevant factors? As I will later show, this question is of crucial importance to evaluating Popper's argument about RP. The answer to this question depends on the reasons we have for regarding idealizations as more true than false. In what follows, then, I return to Nagel's argument to this effect.

[14] For example, Davis Baird notes that in the mid 18th century, the prevailing physical theory about waterwheels did not consider the question of whether the wheel was an overshot waterwheel (water flowing from above) or an undershot wheel (water flowing from below) as relevant to calculating its efficiency. Only when John Smeaton built a working material model of a waterwheel, did it become apparent that this was a causally relevant factor (Baird 2004, 29-32).

Nagel argues that unlike abstractions, idealizations do not fail to provide an exhaustive description of some phenomenon. Unlike assumptions that are false on the evidence (namely, Nagel's second subgroup of unrealistic assumptions), which are literally false, Nagel claims that idealizations are not 'literally false of anything'. Rather, idealizations describe 'pure cases' or 'ideal types' that hold under 'purified conditions' that do not actually exist in reality (Nagel 1963, 215). However, in order to claim this, a clear distinction needs to be drawn between idealizations and mere claims that are false on the evidence. The problem seems to be that interpreted as claims about the world, idealizations are also, strictly speaking, false on the evidence. When interpreted not as claims about the world, claims that are false on the evidence are also not false, since their truth value is determined by their correspondence to the world.

A solution to this problem may be found in Nagel's third group of theoretical assumptions, the set of correspondence rules. These rules establish the way in which theoretical terms refer to phenomena in the world, or the way in which phenomena are to be interpreted in light of the theory. Once we apply these rules, we can make idealizations correspond to the world in a way that enables us to judge their truth value. However, it may be argued that by doing so we have not established the truth value of the idealizations, but only the truth value of another set of statements that contains only non-theoretical terms, which we got by applying the correspondence rules to the idealizations. Nagel's reply to this objection is that it relies on a misguided naive assumption that theoretical terms can be entirely replaced by non-theoretical terms. Since we cannot completely eliminate theoretical terms from our statements about the world, correspondence rules tell us how to take into consideration factors that prevent idealized cases from manifesting themselves in a pure form in the world when we apply idealized statements to the world (Nagel 1963, 215-216).

One way to interpret Nagel is as claiming that idealizations tell the truth about counterfactual cases. Nagel writes: 'discrepancies between what is asserted for the pure case and what actually happens can be attributed to the influence of factors not mentioned in the law' (Nagel 1963, 216). It may be argued that if these factors did not exist, the law would correctly describe reality. For example, while ideal objects to which theoretical terms such as 'vacuum' (Nagel 1963, 216) and 'perfectly divisible and homogeneous commodities' (Nagel 1963, 215) refer do not exist, if they existed, the model would perfectly describe their behaviour.

This claim is not entirely unpersuasive. However, one may worry that this claim is begging the question: From the accurate predictions of the model in real-life cases it infers that the model accurately describes the behaviour of ideal pure cases, while this is exactly what is in dispute. Second, some (but not all) ideal cases, such as Newtonian point-masses, cannot exist in reality without violating other assumptions that are part of the model. If we assume a false image of the world in the first place, it is not clear in what sense the idealizations that are part of it can be counterfactually true.

In light of these worries, one may argue that idealizations describe what happens in real life cases that are close enough to, or the limiting cases of pure cases. Every model gives us a description of the world with respect to some parameters referring to some measurable magnitudes and at some level of precision. When measuring these magnitudes, the level of precision is determined by computational, cognitive, technological and physical limitations, and practical concerns stemming from the use to which we put the model. When we apply the model to the world, we never describe it with infinite precision. We can often create experimental settings that produce close-enough results to what the model predicts for the ideal case, given our desired level of precision. For example, we can suck the air out of a tube so

that the system's behaviour will be the same, up to a tolerable range of error, to that predicted for vacuum. In fact, experimental settings in modern science often aim at doing just that (cf. Ben-Menahem 1988, 169-170).

However, as John D. Norton points out, some idealizations approximate their limiting cases, while others, which are also common in science, do not (Norton 2008, 796).[15] Discussing Norton's example and other possible objections exceeds the scope of this paper. What I hope I have managed to establish is that there is a prima facie plausible case for arguing that abstractions and idealizations are qualitatively different from merely false empirical claims. Crudely speaking, they are hybrids – neither strictly speaking true nor totally false. They capture the significant parts of reality with respect to accurately describing certain aspects of the behaviour of the target system, while they neglect and misrepresent other less or not important aspects.

Since Popper presents RP as an 'oversimplification' and draws an explicit parallel between it and abstractions and idealizations in the physical sciences, in the next section I will argue that rather than expressing a statistical law or an a priori methodological principle, RP captures what Popper views as the important parts of reality with respect to explaining human behaviour, while neglecting what he views as less or not important.

4 Popperian Idealization: A New Interpretation of RP

So far I discussed Popper's analogy between RP and abstractions and idealizations in physics. I presented prima facie reasons to regard abstractions and idealizations as qualitatively different from other claims and as more true than false. I argued that these reasons might underpin Popper's notion of RP. I distinguished between three types of irrelevance that are shielded

[15] Norton describes a model in Newtonian mechanics in which forces in a perfectly rigid beam are statically indeterminate, whereas if the beam has any elasticity, as small as it may be, these forces are statically determinate.

away by abstractions and idealizations: domain-irrelevance, causal irrelevance, and concomitant irrelevance. This enables me now to present a new interpretation of RP, which I will critically evaluate in the next sections.

How can RP be characterized in terms of abstractions and idealizations? From the way Popper characterizes RP, it seems that it is first and foremost an idealization. According to Popper, models in the social sciences are and ought to be constructed on the basis of the assumptions of "pure rationality" on the part of the relevant individual actors. Popper argues that the deviation of actors' actual behaviour from the behaviour that is predicted by the models that make these assumptions is due to the influence of interfering factors that are not stated in the model, such as psychological factors – worries, biases, and the like – that cloud their thought (Popper [1957] 1961, 141). That is, the word "pure" should be interpreted as "uncontaminated", namely as referring to an ideal agent who correctly figures out the objective situation in which she is found and the best course of action in that situation, and whose thought is not limited, obstructed or distracted in any way. RP is therefore first and foremost an idealization.

It is possible that such thought-clouding factors are relatively rare, and in this case the idealization will describe what happens in most cases, as the statistical interpretation of RP suggests. But it is not necessarily so. Such interrupting factors may also be quite common. As will show in the last section, in some cases, we may still construct a successful model employing RP by systematically adding to it corrections that take into account the effect of such interrupting factors. Assumptions about the frequency of cases in which people's behaviour deviates from their expected behaviour in pure conditions of complete rationality are therefore not inherently part of RP.

What about assumptions about the mental states of the actors in the model? As mentioned above, Popper deems them irrelevant to the construction of the model. In my view, it is useful in this context to distinguish between two types of irrelevant facts about the actors' states of mind. Recall that Popper argues that actors' general knowledge and beliefs about things that are not part of the analyzed situation are not relevant to the model. For example, in the case of Richard's crossing the road, Popper argues that Richard's knowledge of Sanskrit texts is irrelevant to the situation.

Note that an actor's general knowledge is domain-relevant to the situation. This is because the actor is part of the situation, and a complete description of that actor, hence a complete description of the situation, includes his or her general knowledge and beliefs. However, an actor's general knowledge and beliefs or at least much of them are arguably causally irrelevant to the situation. Richard's general knowledge of Sanskrit texts, for example, is causally irrelevant to the situation, because it does not causally affect, or so Popper seems to claim, the aspect of Richard's behaviour that we want to describe, namely his crossing the road.

What about evidence about actors' psychological motivations? As you recall, Popper deems it irrelevant to the model. Popper argues that evidence obtained from questionnaires about businesspersons' psychological motivations is irrelevant to an economic model explaining their behaviour. Furthermore, Popper rejects an instrumentalist interpretation of such a model as a merely useful tool for prediction. Rather, he defends a realist interpretation of such an economic model, even though it lacks any reference to the businesspersons' psychological motivations (Popper 1994, 182). At the same time, Popper acknowledges that in some cases, such as the frustrated driver, various psychological mechanisms may cause people to act irrationally (Popper 1994, 172). From this we can conclude that according to Popper, evidence about people's psychological motivations may be causally relevant to the description

of people's actual behaviour. Furthermore, since as I have shown, RP does not involve an assumption about the frequency of such cases, such evidence may be causally relevant to many cases. In what sense, then, is psychological evidence irrelevant to the model?

I argue that for Popper, psychological evidence is concomitantly irrelevant to models of social situations. Namely, actors' psychological motivations, which may causally affect their behaviour, are screened off by the idealizing assumption that people act in accordance with the situation. In the ideal gas model, the idealizing assumption that there are no intermolecular forces between the gas molecules screens off the inner-structure of the molecules. It allows us not to worry about what is going on inside the molecules. Similarly, in situational analysis, the assumption that agents act rationally, i.e. in accordance with the situation, screens off their psychological motivations. As we assume that agents act in pure rationality, we need not worry about what is going on inside their heads. This is not because what is going on inside their heads does not causally affect their actual behaviour. On the contrary, it does. Rather, our ability to keep silent about the agents' mental states is a consequence of an idealization that exists in our model.

To sum up, in this section I have presented and defended a new interpretation of Popper's rationality principle, which explains why Popper claims that it allows scientists to construct approximately true models. RP consists of the following three assumptions:

RP1 – Agents act in "pure rationality", namely according to the most adequate course of action to achieve the goals that are part of the objective description of the situation. This is the way agents would act if their thought were not limited, distracted, or otherwise obstructed. (An idealization.)

RP2 – Descriptions of agents' general knowledge and beliefs are not part of the model. (An abstraction away from causally-irrelevant factors.)

RP3 – Descriptions of agents' general mental states and psychological motivations with regard to their actions are not part of the model. (An abstraction away from concomitantly-irrelevant factors that are screened off by RP1.)

My formulation of RP improves our understanding the role of RP in the social sciences. It still, however, contains terms such as 'the objective description of the situation', which require further unpacking. In the next section, I will unpack these terms and critically discuss the privileged status Popper ascribes to RP as a golden foundation of the social sciences.

5 RP1, RP2, RP3 and the Problem of Seemingly Irrational Behaviour

Popper maintains that RP is a golden foundation of all the social sciences. He believes it has a privileged status as the animating law for all the models in the social sciences. Popper's defence of the privileged status of RP is twofold. First, he provides a general argument for why always adopting RP is a wise methodological policy. Second, he discusses various types of social situations that seem incompatible with RP, and argues that they are not so.

Popper's general argument in defence of the privileged status of RP is fourfold. In a short and somewhat cryptic paragraph, Popper lists four reasons why always employing RP is methodologically wise. First, a model is far more informative and better testable with RP. Popper claims that we already know that RP is, strictly speaking, false, or in other words, RP is not the component of the model being put to the test. Second, Popper claims that RP is sufficiently near to the truth. Thus, it is not likely to be responsible for a breakdown of the model. Rather, it seems that Popper believes that if the model is falsified, an incorrect or inaccurate analysis of the situation will be responsible for that. As I have argued, Popper believes RP is sufficiently near to the truth in the sense that it captures the aspects of human behaviour that are most relevant to its explanation. Third, Popper states that an attempt to replace RP will lead to complete arbitrariness in our model building. Fourth, Popper observes

that most social scientific theories include RP. Therefore, when deciding between competing models, RP itself will usually not be challenged (Popper 1994, 178).[16]

None of these arguments, though, definitely ensures the preservation of RP. While it may be improbable that RP will be responsible for the failure of a theory, it is still possible. While it creates models that are sufficiently near to the truth, it may be found out that they are not sufficiently near to the truth or that other models that do not employ RP are as near or nearer to the truth. While RP distinguishes relevant from irrelevant factors, other distinguishing criteria may be suggested and proven to be successful. While most current theories employ RP, it may not be so in the future. Since according to Popper, every theory may be replaced by a better theory in the future, and every false empirical claim in a theory may be taken out of it, this argument does not seem to guarantee the preservation of RP in every theory.

In order to better defend the secure status of RP, Popper needs to show that it applies to all types of social situations, even those that seem to reject it. Such cases are typically cases in which agents seem to act irrationally. Only then will we have good reasons to believe that giving up RP is unwise. Popper identifies two types of cases in which people allegedly do not act rationally and argues that they can still be successfully analyzed using RP. The first is the case of the incompetent war leader: A war leader makes mistakes in managing the war, which cause his army to lose, while his actions seem to the objective outside observer as inadequate in the situation. The second case is that of the neurotic person. Popper argues that a rational analysis in both this cases is possible. What we need to do is to subjectivize the analysis of the situation to the agents' points of view, and see that their actions are adequate to the way they see the situation. The incompetent war leader errs because of his limited experience, limited or

[16] For a different interpretation of this fourfold argument that rests on the statistical interpretation of RP see Lagueux (1993) at 174-476.

overblown aims, or limited or overexcited imagination. Interestingly, Popper argues that Freud's theory of the typical origin of neurosis is perfectly compatible with RP as well, because Freud explains neurosis as a rational attitude for a child to develop in an early stage of his or her life when he or she is unable to cope with the situation in a better way (Popper 1994, 178-179).

Popper emphasizes that when subjectivizing the situation to the actor's point of view, the scientist's aim is not to achieve understanding by getting into the actor's mind. Rather, in a manner consistent with my interpretation of RP, he states that the aim is 'to produce an idealized and reasoned reconstruction of it [i.e. the situation], omitting inessential elements and perhaps augmenting it', where the situational analysis is a conjecture that needs to be tested (Popper [1972] 1979, 188; emphasis added). Namely, Popper insists that the description of the situation and its analysis remain objective.

Popper distinguishes between three senses of the situation: (1) this situation as it really was; (2) the situation as the agent actually saw it; (3) the situation as the agent could (or ought to) have seen it within the objective situation – this is the sense of the situation that corresponds to the "pure rationality" requirement of RP1 on my interpretation or RP, namely the situation as an ideal agent whose thought is not clouded would have seen it. Note that (2) and (3) are part of (1) as they are part of the complete objective situation. Popper argues that explanations of failure (such as the incompetent war leader) explore the difference between (1) and (2). In cases where there is a difference between (2) and (3), we will say that the agent acts irrationally (Popper 1994, 183).

This analysis is supposed to show that RP is applicable to cases in which people do not act rationally with respect to the objective situation as seen by an objective observer. However, several concerns arise. First, Popper acknowledges that when there is a discrepancy between

(2) and (3), people simply act irrationally. Suppose we want to explain these cases, how can RP apply to them? Popper's answer seems to be the one he gives to the case of the neurotic person. We invoke RP to explain why it was rational to adopt certain patterns of behaviour, such as neurosis, in certain situations, like childhood. Nevertheless, this explanation does not explain why as an adult, the neurotic person fails to see the situation correctly, and act in accordance with it, but rather sticks to his or her childhood neurosis.

In response to this problem, one can simply admit that such situations cannot be explained by the social sciences. However, this seems implausible in cases in which people systematically act irrationally. The existence of a pattern of irrational behaviour suggests the existence of an explanation as well. I will return to this point in the end of this section.

If we do want to give social-scientific explanation of phenomena such as neurotic behaviour in adults, there seems to be two possible ways. One way is to revert to some psychological laws, such as a law that states that people find it difficult to lose habits they develop in childhood. However, this explanation resorts to psychologism, more precisely psychological nomologism, which is exactly what Popper wants to avoid. Popper is aware of this problem, and admits that certain human psychological traits such as susceptibility to propaganda may sometime cause people to deviate from rational behaviour, in the sense just discussed. He further admits that certain basic psychological motives such as self preservation or avoiding pain may explain the behaviour of people. Nevertheless, he claims that such psychological parts of the explanation are 'trivial'. (Popper [1961] 1966, 96-97).

However, as I will now illustrate using Daniel Kahneman and Amos Tversky's prospect theory, such psychological components are not necessarily trivial or marginal. To the extent that they are not, it seems that Popper must concede that they cannot be explained using situational analysis. This stands at odds with Popper's claim that psychology is and ought to be

reducible to situational analysis rather than vice versa (Popper [1957] 1961, 142; Popper [1961] 1966, 97-98). At the same time, I will argue that while psychological nomologism is necessary in explanations of deviations from rationality, RP may still play an indispensable role in explanation of human behaviour as well and this is without resorting to the second form of psychologism, i.e. mentalism.

Kahneman and Tversky's prospect theory (Kahneman & Tversky 1979) is an alternative to the expected utility theory. According to the expected utility theory, when making decisions under risk, people tend to maximize their utility. This theory was largely accepted as a descriptive model of economic behaviour. In a series of experiments, Kahneman and Tversky found that people do not act in order to maximize their utility. Rather, they tend to avert risks in choices involving sure gains, and to seek risk in choices involving sure loses. This tendency, which they call 'the certainty effect', creates inconsistent preferences with respect to the ones predicted by the expected utility theory.

The following reconstructed example may clear the differences between prospect theory and the expected utility theory. Suppose you have two choices:

(a) You receive $200. Would you prefer (a1) to get additional $100 or (a2) to get additional $200 with a 0.5 probability?

(b) You receive $400. Would you prefer (b1) to return $100 or (b2) to return $200 with a 0.5 probability?

Note that options (a1) and (a2) are equivalent to (b1) and (b2), respectively, in terms of the final utility.[17] Therefore, utility theory expects that if a subject chooses (a1) she will also choose (b1). However, Kahneman and Tversky found that subjects who choose (a1) tend to choose (b2). This is consistent with the certainty effect of prospect theory, according to which

[17] If you choose (a1) or (b1) you end up with $300. If you choose (a2) or (b2) you will end up either with $400 or $200, where the two outcomes have an equal chance.

people tend to avoid risks in choices involving sure gains, and prefer risks in choices involving sure loses.

What is the place of RP in Kahneman and Tversky's prospect theory? As first blush, prospect theory refutes RP – it shows that people systematically fail to act according to the objective situation. This seems to be a general problem for assumptions of rationality in classical economics.[18] Merrilee H. Salmon argues that in light of Kahneman and Tversky's findings, a principle such RP is either hopelessly false or empty of any empirical content:

> It seems that with enough ingenuity we can always attribute an "appropriate" set of beliefs and desires to make any action count as rational. However outlandish an action, it is always open to us to say, "It seemed (to the agent) a good idea at the time." Such a possibility weakens the claim that "Agents always perform those actions with greatest desirability" has any empirical content. In the words of Karl Popper, the alleged law is unfalsifiable […] it seems that empirical content for the principle that agents always perform those actions with greatest expected desirability can be salvaged only at the price of rendering the principle false. In either case, the principle is not a suitable law to ground scientific explanation (Salmon 1992, 414-415).

Salmon argues that since explanatory covering laws need to have empirical content, a principle like RP cannot be such a law. As we have seen, this stands at odds with Popper's claim that RP is the animating law of all explanatory models in the social sciences.

As mention above, Popper's general strategy for dealing with discrepancies between agents' actual seemingly irrational behaviour and their expected behaviour is reconstructing

[18] Theories in economics typically employ notions of instrumental rationality. Under instrumental rationality, a subject is attributed beliefs and ends, and is considered to act rationally if she acts in the best way to achieve her ends according to her beliefs. Formally speaking, a subject has a utility function that consistently assigns different utilities to the subject's different aims. The subject acts rationally when she maximizes her expected utility. Instrumental rationality does not pass judgment on the beliefs and aims themselves. While instrumental rationality per se does not attributes aims and beliefs to the subject, economic theory must do so in order to explain human behaviour. It typically assumes that agents seek to maximize profit and reduce risk (Weirich 2004, 380-381). In spite of Popper's references to economics, in my view, RP should not necessarily be equated with the notion of instrumental rationality that is employed in contemporary economics. This is for two reasons. First, instrumental rationality, as such, requires that the agent act consistently with respect to her preferences, but does not say anything about the rationality of the preferences themselves. By contrast, it seems that applying Popper's RP may involve passing such judgment on the agent's preferences in some cases. Second, it is not clear that the notion of utility is an integral part of the objective description of any social situation. For example, in the case of Richard's crossing the road, an appeal to utility may seem artificially imposed on the situation, and indeed, Popper does not list it as part of its elements.

the situation from the agents' limited and perhaps distorted perspective, and then explaining the rationality of their behaviour. Salmon worries that such a move may be no more than an ad hoc attempt to save the appearances. Is she right? I will argue that such a move, carefully done, does not empty RP of empirical content, but that in some cases it comes at the cost of abandoning RP2 and RP3. Before I do that, let us briefly examine an additional worry raised by Peter Winch, which leads to the same conclusion.

Winch distinguishes between aims that are intrinsic to the situation and aims that are extrinsic to it. For example, when I am playing chess and I am asked what my aim is, I may reply that my aim is to capture the opponent's king (intrinsic aim) or to impress my boss, who thinks highly of chess players (extrinsic aim). A description of the frustrated driver's desire to express his protest against the poor administration of the city is another example of having an extrinsic aim. Winch argues that Popper's philosophy of social sciences does not distinguish between the two types of aims, which derogates from its explanatory power (Winch 1974, 895-896).[19] In the present context of RP, the introduction of the actor's subjective point of view divorces the rules of the situation from its aims. This causes an indeterminacy of the outcome. Suppose I am playing chess with my boss and my intention is to suck up to him. Then I will purposely lose the game. Suppose, on the other hand, that I am in a chess tournament, then I will play in order to beat my opponent. In both cases, I am following the same rules of chess. However, my moves and probably the outcome of the situation will be very different.

Once again, we see that all seemingly irrational behaviour can be reinterpreted as rational if we postulate an appropriate extrinsic aim. This puts RP at the danger of becoming

[19] Popper replies to this criticism by saying that his philosophy of social sciences does distinguish intrinsic from extrinsic aims. He also makes reference to various places in which he claims to make this distinction (Popper 1974, 1167-1168). However, as far as I know, nowhere does Popper address the application of this distinction, inasmuch as he makes it, to his view about the subjectivization of RP, which is the central issue here.

empty of empirical content and unfalsifiable. As Popper strongly wishes to avoid this consequence, a clear demarcation criterion needs to be established between legitimate interpretations of the situation, which postulate legitimate extrinsic aims, and illegitimate interpretations, which postulate illegitimate extrinsic aims. To do that, we need to know if the way in which we describe the situation through the actors' eyes is actually the way they see it. In order to know this, it seems unavoidable to ask the actors how they see the situation and compare their answers with the way we think they see the situation and what we think that their aims are. One way to ask the actors is distributing questionnaires about their motivations.

Of course, we need not take actors' answers to questionnaires at their face value, and we are allowed to infer other beliefs, motivations and aims from their answers or other evidence.[20] The point is that evidence about their beliefs and aims is relevant. If we want to explain discrepancies without being charged of merely saving the appearances, we must take into account the actors extrinsic aims, motives and beliefs, which means that we cannot abstract away from the actors' general beliefs (RP2) or their psychological motivations (RP3). Therefore, pace Popper, at least in some cases, if RP is to retain its empirical content, evidence about agents' psychological motivations and general beliefs is necessary, and we are forced to violate RP2 and RP3.

To sum up the discussion so far, the problem with introducing subjectivity as a way of accounting for seemingly irrational behaviour is the following. On the one hand, Popper acknowledges that we need to introduce an analysis of the situation from the actors' subjective point of view – otherwise we will not be able to explain situations in which actors do not act in accordance with the objective situation. On the other hand, the introduction of the actors'

[20] The fact that beliefs and mental states are unobservable is not in itself a problem for the falsifiability of situational analyses that employ RP, at least no more that the references to unobservable entities in physical theories.

subjective point of view requires – at least in some cases – that we not abstract away from the actors' general beliefs and psychological motivations. This leads to abandoning RP2 and RP3, which Popper regards as central to RP.

While abandoning RP2 and RP3 cuts us clear of the danger of RP losing its empirical content, it does not necessarily preserve its privileged status. It may be empirically discovered that in some situations, people do not act rationally, even according to the way they see the situation. In addition, abandoning RP2 and RP3 makes RP much less attractive. RP is attractive exactly because it helps us simplify a complex reality and construct approximately true models. Without RP2 and RP3, this stops being the case.

By giving up RP2 and RP3, then, we may be throwing out the baby with the bathwater. Is there another way to preserve RP that avoids the problem just discussed? In the next section I will suggest such a way and illustrate it using Kahneman and Tversky's prospect theory.

6 Resolving the Problem: RP as a Systematically De-Idealizable Idealization

As we have seen, Kahneman and Tversky found that when faced with equivalent dilemmas in terms of final utility, people do not consistently act to maximize their expected utility, namely they do not act according to the objective situation. How should proponents of RP react to these findings? So far, I have presented two options. The first is to declare RP false and abandon it. The second is to reconstruct the situation from the agents' point of view and try to rationally account for their behaviour. The second option necessitates gathering empirical evidence about agents' beliefs and aims (to avoid the charge of merely saving the appearances), and does not guarantee success. Both options are unattractive to proponents of RP. I would like now to suggest a third option, according to which RP is an idealization which may be systematically de-idealized.

Note that without an assumption that people want – in some way – to maximize their gains, that they prefer gains over loses, that they like having more money than less, or something like that, prospect theory is a non-starter. I suggest viewing this minimal assumption about people's general like of richness and dislike of poorness as corresponding to the RP1 component of RP, namely, the assumption that agents act in "pure rationality". This assumption may be minimal but is a necessary starting point. This is consistent with Popper's claim that RP is 'an almost empty principle', in the sense that it assumes very little about the actors, but it is necessary in order to 'animate' the model. In addition, this explains why Popper believes that the absence of RP means total arbitrariness (Popper 1994, 178) – if we do not assume that people prefer to be rich than poor, or something of this sort, their decisions and actions become unintelligible.

What about the discrepancies between what is expected in conditions of pure rationality and people's actual behaviour? Here comes to play the notion of systematic de-idealization. Margaret Morrison presents the model of the simple pendulum as a paradigmatic case of model-using in science. We start with a highly abstract version of the real thing, a description of an idealized pendulum. We then start systematically adding corrections to it, such as the resistance of air, and the mass of the cord, in order to make the model more and more realistic and its predictions more and more accurate. Morrison argues that our ability to systematically add such corrections, or de-idealizations, stems from the robustness of our background theory (Morrison 1999, 48-53).

I suggest that the same analysis is applicable to Kahneman and Tversky's prospect theory, and probably other theories in the social sciences. First we start with the expected utility theory, which assumes a highly ideal version of human behaviour. In this theory the value, or the utility, one ascribes to his or her choice is symmetrical in terms of gains and

losses, and the marginal utility decreases with the increase of gains. What prospect theory does is to systematically correct this in light of the empirical findings. The function graph in figure 1 illustrates this.

[Insert figure 1 about here]

Figure 1 (Kahneman & Tversky 1979, 279)

The X axis represents one's gains or losses. The Y axis represents the utility. According to the utility theory, if we drew the objective utility as a function of the gains/losses, we would expect it to increase (decrease) with the gains (losses), and for its rate of increase (decrease) to decline with the gains (losses). This expectation represents the assumption that the marginal utility is diminishing (Rosenberg 1988, 69). Moreover, we would expect the objective utility to be symmetrical with respect to the main diagonal, namely for the marginal utility to increase and decrease at the same rate.

Note, however, that the curve on the left side is steeper, which represents the fact that people ascribe a larger absolute subjective value to a given loss than to the equivalent gain. I suggest seeing the deviance of this graph from the graph expected from the utility theory as a de-idealization. While the utility theory gives us an idea about how the graph should generally look like, the graph that represents the empirical finding has the same general shape, except for the asymmetry with respect to the main diagonal. This function is therefore a systematic de-idealization of the complete rationality assumption (RP1).

This analysis of the way RP is employed in prospect theory is consistent with Popper's account of it in two ways. First, Popper says that RP acts as a 'zero-method'. Popper says that the people's behaviour in conditions of pure rationality serves as a kind of zero coordinates, from which people's actual behaviour may deviate (Popper [1957] 1961, 141). The above example of systematic de-idealization shows how this is carried out in practice. The deviation

of reality from the zero coordinates is not measured by the number of people who deviate from pure rational behaviour, as the statistical interpretation of RP would seem to suggest. Rather, it refers to our ability to systematically correct the predictions of the model, so that they will correspond to the values measured in reality (cf. Jarvie 2001, 133-134).

Second, the introduction of the de-idealization by itself does not involve assumptions about people's psychological motivations or general beliefs. While the introduction of the subjective point of view as a de-idealization may introduce one form of psychologism, i.e. psychological nomologism (which is, as you recall, an appeal to psychological laws) to explain actors' deviation from what is expected under the assumption of pure rationality, it does not resort into the second form of psychologism – mentalism, which is an explanatory appeal to people's mental states. As such, it violates neither RP2 nor RP3, and can still be seen as an application of RP to the model.

Such application and de-idealization of RP is not unique to prospect theory. Assigning a higher subjective value to a gain than the equivalent loss is not the only way in which agents deviate from pure rationality. As William Gorton notes, experimental psychologists have documented many ways in which agents form beliefs and make decisions in an irrational yet predictable way. Such ways are broadly dividable into two camps. The first camp is of "cold" cognitive biases, in which agents reason in a faulty manner due to using incorrect rules of probability or incorrect evaluation of the evidence. The second is of "hot" cognitive biases, such as wishful thinking, in which agents' reasoning is affected by their goals or desires.[21] Gorton demonstrates how Jon Elster's (1993) theory of the formation of political revolutions may be seen a Popperian situation analysis that first employs RP, and then systematically corrects – or de-idealize, on my account – the pure rationality assumption by systematically adding different

[21] See Kunda (1990) and Plous (1993) for detailed overviews.

psychological mechanisms of biased reasoning to the model until it correctly accounts for the observed reality (Gorton 2006, 110-119). The rich body of research from experimental psychology on human faulty reasoning enables, in principle, the application of such de-idealizations to different models.

However, this analysis also exposes the weaknesses of RP as an always-privileged methodological principle. What happens if we want to explain the de-idealizations that we have introduced into the model? If we want to explain, for example, why people hate to lose more than they like to gain, we can resort to various cognitive, psychological, and perhaps evolutionary explanations. However, as I have argued, such explanations of the de-idealizations themselves cannot be based on RP, as they aim to account for irrational behaviour.

As Morrison argues, our ability to systematically correct our abstract models stems from the robustness of our background theories. These theories cannot be based on RP. Therefore although the prospect theory example shows that RP may play a central role in social scientific explanations, it also shows that it has inherent limitations, which prevent it from serving as the golden foundation of all the social sciences, as Popper hoped it to be.

Conclusion

Popper presents an intricate view of the role of models in science not merely as helpful heuristic devices, but as theoretical explanatory devices as well. Popper equates explanatory models in the social sciences with analyses of social situations. What animates these models, in his view, is RP, which is the assumption that people act in accordance with the objective situation.

In this paper I have presented a new interpretation of Popper's RP. I have argued that Popper's RP consists of an idealization according to which people act in pure rationality, and two abstractions. The first abstraction omits agents' general beliefs, which Popper deems what

I define as causally irrelevant to the model. The second abstraction omits agents' psychological motivations and mental states. These factors are what I define as concomitantly-irrelevant to the model, as they are screened off by the main idealization of RP.

Having presented my interpretation, I have critically examined the privileged status Popper ascribes to RP in the social sciences. I have argued that RP may play an important role as a central idealization in social scientific models. Using Kahneman and Tversky's prospect theory, I have illustrated that as an idealization, RP can be systematically de-idealized to better account for the empirical data without introducing agents' beliefs and psychological motivations into the model. However, I have also shown that if we want to explain these deviations from the model, we cannot invoke RP to do that.

References

Agassi, Joseph. 1960. Methodological individualism. The British Journal of Sociology 11(3): 244-270.

Baird, Davis. 2004. Thing knowledge: A philosophy of scientific instruments. Berkeley: University of California Press.

Ben-Menahem, Yemima. 1988. Models of science: Fictions or idealizations? Science in Context 2(1): 163-175.

Cartwright, Nancy. 1983. How the laws of physics lie. New York: Oxford University Press.

Cartwright, Nancy. 1989. Nature's capacities and their measurement. New York: Oxford University Press.

Chakravartty, Anjan. 2007. A metaphysics for scientific realism: Knowing the unobservable. Cambridge: Cambridge University Press.

Elster, Jon. 1993. *Political psychology*. Cambridge: Cambridge University Press.

Feyerabend, Paul. 1981. Realism, rationalism and scientific method. Cambridge: Cambridge University Press.

Gorton, William A. 2006. Karl Popper and the social sciences. Albany, NY: SUNY Press.

Hempel, Carl G. & Paul Oppenheim. 1948. Studies in the logic of explanation. Philosophy of Science 15(2): 135-175.

Jarvie, Ian C. 2001. The republic of science: The emergence of Popper's social view of science 1935-1945. Amsterdam: Rodopi.

Jones, Martin R. 2005. Idealization and abstraction: A framework. In Idealization XII: Correcting the model. Idealization and abstraction in the sciences, edited by Martin R. Jones and Nancy Cartwright, pp. 173-217. Amsterdam: Rodopi.

Kahneman, Daniel & Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. Econometrica 47(2): 263-292.

Keuth, Herbert. 2005. The philosophy of Karl Popper. Cambridge: Cambridge University Press.

Koertge, Noretta. 1979. The methodological status of Popper's rationality principle. Theory and Decision 10(1): 83-95.

Kunda, Ziva. 1990. The case for motivated reasoning. Psychological Bulletin 108(3): 480-498.

Lagueux, Maurice. 1993. Popper and the rationality principle. Philosophy of the Social Sciences 23(4): 468-480.

Miller, David. 2006. Out of error: Further essays on critical rationalism. Aldershot: Ashgate.

Morrison, Margaret. 1999. Models as autonomous agents. In Models as mediators: Perspectives on natural and social science, edited by Mary S. Morgan and Margaret Morrison, pp. 38-65. Cambridge: Cambridge University Press.

Nagel, Ernest. 1963. Assumptions in economic theory. The American Economic Review 53(2): 211-219.

Norton, John D. 2008. The dome: An unexpectedly simple failure of determinism. Philosophy of Science 75(5): 786–798.

Notturno, Mark A. 1998. Truth, rationality and the situation. Philosophy of the Social Sciences 28(3): 400-421.

Plous, Scott. 1993. The psychology of judgment and decision making. New York: McGraw-Hill.

Popper, Karl R. [1957] 1961. The poverty of historicism, 3rd ed. London: Routledge.

Popper, Karl R. [1961] 1966. The open society and its enemies, Vol. II, 5th ed. Princeton: Princeton University Press.

Popper, Karl R. 1967. La rationalité et le statut du principe de rationalité. In Les Fondements philosophique des systèmes économiques, edited by E. M. Classen, pp. 142 150. Paris: Payot.

Popper, Karl R. 1974. Winch on institutions and the open society. In The philosophy of Karl Popper: Book II, edited by Arthur P. Schilpp, pp. 1165-1172. La Salle, IL: Open Court.

Popper, Karl R. [1972] 1979. Objective knowledge: An evolutionary approach, Rev. ed. Oxford: Oxford University Press.

Popper, Karl R. 1985. The rationality principle. In Popper selections, edited by David Miller, pp. 357-365. Princeton: Princeton University Press.

Popper, Karl R. 1994. Models, instruments, and truth: The status of the rationality principle in the social sciences. In The myth of the framework: In defence of science and rationality, edited by Mark A. Notturno, pp. 154-184. London: Routledge.

Rosenberg, Alexander. 1988. Philosophy of social science. Boulder, CO: Westview Press.

Salmon, Merrilee H. 1992. Philosophy of the social science. In Introduction to the philosophy of science: A text by members of the department of the history and philosophy of science of the University of Pittsburgh, edited by Merrilee H. Salmon et al, pp. 404-425. Englewood Cliffs, NJ: Prentice Hall.

Schaffner, Kenneth F. 1993. Discovery and explanation in biology and medicine. Chicago: University of Chicago Press.

Teller, Paul. 2001. Twilight of the perfect model model. Erkenntnis 55: 393-415.

Weirich, Paul. 2004. Economic rationality. In The Oxford handbook of rationality, edited by Alfred R. Mele and Piers Rawling, pp. 380-398. Oxford: Oxford University Press.

Winch, Peter. 1958. The idea of a social science and its relation to philosophy. London: Routledge.

Winch, Peter. 1974. Popper and scientific method in the social sciences. In The philosophy of Karl Popper: Book II, edited by Arthur P. Schilpp, pp. 889-904. La Salle, IL: Open Court.