THE UNIVERSITY
of EDINBURGH

TOWARDS A THEORY OF ADAPTIVE RATIONALITY?

Andrea Polonioli

PhD in Philosophy

The University of Edinburgh

2015

Declaration of Own Work

I hereby declare that the present thesis is entirely my own work, except where otherwise indicated by means of quotation, reference and acknowledgement. The work has not been submitted for any other degree or professional qualification.

Signed: Date:

Table of Contents

Acknowledgements

I would like to extend my gratitude to everyone who helped me in one way or another to complete this dissertation. My greatest debt is to my supervisors at the University of Edinburgh, Tillmann Vierkant and Michela Massimi, who have shepherded me through various aspects of this dissertation. They have been excellent advisors and were a constant source of inspiration throughout the writing of my PhD thesis. They have provided ample professional advice and invaluable feedback on my work. And most important of all, their support allowed me to genuinely enjoy working on my thesis.

Yet, as often happens, one's PhD is the result of different times and places, and since the questions that motivate this thesis originated from my time as a Masters student at the San Raffaele University in Milan, I also wish to thank Francesco Guala, who supervised my Masters thesis, for his detailed and insightful comments and very stimulating conversations.

Part of this work was carried out also during my visits to the Centre for the Philosophy of Natural and Social Sciences at the London School of Economics and to the Department of History and Philosophy of Science at the University of Pittsburgh. I am very much indebted to all the people who made these visits possible and so useful.

For their conversation and correspondence regarding various topics addressed in this thesis I am particularly indebted to Wendy Johnson, Campbell Brown, Matteo Colombo, Till Grüne Yanoff, Lars Penke, and Armin Schulz.

I also want to thank the College of Humanities and Social Sciences at the University of Edinburgh and the Royal Institute of Philosophy for generously funding my research. Without their support my work on this thesis would not have been possible.

I would like to thank Battista and Valentina, my family, for their support and patience. Throughout the writing process, they unfailingly supported me in coping with my absenteeism and tolerating my dissertation-related stress.

Material from this thesis has been presented in Santa Cruz, Charleston, Stockholm, Bristol, Exeter, Sheffield, Trento, Berlin, and Rotterdam. I am grateful to the audiences at all of these events for their stimulating and invaluable feedback.

I should also acknowledge that material from Chapter One has been published in *Mind & Society*, material from Chapter Two in *Biological Theory*, material from Chapters Three and Seven in *Philosophy of the Social Sciences*, material from Chapter Four in *Frontiers in Psychology*, and material from Chapter Six in *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of the Biological and Biomedical Sciences*.

I humbly dedicate this thesis to Mariana, who made the whole experience a lot more enjoyable.

Abstract

The idea that humans are prone to widespread and systematic biases has dominated the psychological study of thinking and decision-making. The conclusion that has often been drawn is that people are irrational. In recent decades, however, a number of psychologists have started to call into question key claims and findings in research on human biases. In particular, a body of research has come together under the heading of adaptive rationality (henceforth AR). AR theorists argue that people should not be assessed against formal principles of rationality but rather against the goals they entertain. Moreover, AR theorists maintain that the conclusion that people are irrational is unsupported: people are often remarkably successful once assessed against their goals and given the cognitive and external constraints imposed by the environment. The growth of literature around AR is what motivates the present investigation, and assessing the plausibility of the AR challenge to research on human biases is the goal of this thesis. My enquiry analyses several aspects of this suggested turn in the empirical study of rationality and provides one of the first philosophically-informed appraisals of the prospects of AR. First and foremost, my thesis seeks to provide a qualified defence of the AR project. On the one hand, I agree with AR theorists that there is room for a conceptual revolution in the study of thinking and decision-making: while it is commonly argued that behaviour and cognition should be assessed against formal principles of rationality, I stress the importance of assessing behaviour against the goals that people entertain. However, I also contend that AR theorists have hitherto failed to provide compelling evidence in support of their most ambitious and optimistic theses about people's rationality. In particular, I present a great deal of evidence suggesting that people are often

unsuccessful at achieving prudential and epistemic goals and I argue that AR theorists have not made clear how, in light of this evidence, optimistic claims about human rationality could be defended.

0. Introduction

0. 1 The subject matter

What a piece of work is a man, how noble in reason, how infinite in faculties, in form and moving how express and admirable, in action how like an angel, in apprehension how like a god: the beauty of the world, the paragon of animals.
William Shakespeare

We must further observe that while our inferences from experience are frequently fallacious, deduction [...] cannot be erroneous [...]. My reason for saying so is that none of the mistakes which men can make (men, I say, not beasts) are due to faulty inference; they are caused merely by the fact that we found upon a basis of poorly comprehended experiences, or that propositions are posited which are hasty and groundless.
Descartes

Rationality is currently a hot topic in scientific research—and a particularly difficult one too. Aristotle famously defined human beings as rational animals, and even today few people would openly describe themselves as irrational. In fact, to many people the idea that human beings are rational may sound like a platitude. But the assumption that we are rational agents has been popular in scientific domains too—in particular, it has been central to theorizing in the social sciences. In the domain of social policy, for example, the rationality assumption has often been used to support the idea that it is not necessary to protect people against the consequences of their choices.

In recent decades, however, the assumption that people are rational has been empirically tested. And the results of many years of scientific research on human rationality seem to pose a number of challenges to optimistic assumptions about human rationality. More precisely, in the 1970s researchers started to show that

human beings commit systematic and widespread reasoning errors. As a result, the view that our capacities for reasoning and decision-making are severely flawed has become dominant in the psychology of judgment and decision-making. Notably, bleak assessments of human rationality have travelled fast outside the field of decision science and into other disciplines concerned with human behaviour, such as marketing science or behavioural finance, where the view that human beings are irrational now figures prominently in popular handbooks.

Science moves fast, however, and the proposition that people are irrational has recently attracted fierce criticism too. In particular, a new strand of research has come together under the umbrella term "adaptive rationality" (henceforth AR), and this framework has been presented as a radically alternative approach to mainstream empirical research on human rationality. These theorists suggest that scientific research into human rationality has been led astray: the findings suggesting human irrationality that have been reported by researchers in the field of judgement and decision-making should not in fact be seen as indications of human irrationality, but rather as a result of applying the wrong experimental formats and—more importantly—the wrong normative standards. In particular, AR theorists argue that people should not be assessed against norms of logic, probability theory, and decision theory, but rather against the goals they entertain. They also maintain that the conclusion that people are irrational is unsupported: people are often remarkably successful once assessed against their goals and given the cognitive and external constraints imposed by the environment. The AR project has been often presented as attempting to show 'how people are able to achieve intelligence in the real world'

(Todd and Gigerenzer 2012, 3), and provocative book titles such *Adaptive thinking: Rationality in the Real World* (Gigerenzer 1999) or *Ecological Rationality: Intelligence in the Real World* (Todd and Gigerenzer 2012) illustrate quite clearly the general aims of their project. In arguing for such claims, AR theorists attempt to restore optimistic assumptions about human rationality, although they do not do so by restating the so-called "Normative Man" hypothesis, namely the idea that human reasoning and decision making can be roughly modeled by Expected Utility theory (Edwards 1954). Rather, they do so by putting forward a new normative perspective on human rationality.

The growth of this literature and trend is what motivates the present investigation, and assessing the plausibility of this challenge is the very goal of this thesis. My enquiry analyses several aspects of this suggested turn in the empirical study of rationality and provides one of the first philosophically-informed appraisals of the prospects of the challenge from AR.

This thesis aims to provide a qualified defence of AR. On the one hand, I agree with AR theorists that there is room for a conceptual revolution in the study of thinking and decision-making: while it is commonly argued that behaviour and cognition have to be assessed against formal principles of rationality, I stress the importance of assessing behaviour against the goals that people entertain. However, *contra* AR theorists, I also present a great deal of evidence suggesting that people are often unsuccessful at achieving prudential and epistemic goals. I argue that AR theorists have not made clear how, in light of this evidence, optimistic claims about

human rationality could be defended. I therefore conclude that, while AR theorists have pushed the "rationality debate" forward, they have also failed to provide compelling evidence or reasons in support of their most ambitious theses.

Before I begin to articulate my argument in detail, there are a few important tasks I wish to accomplish in this introductory chapter. In sections 0.2 and 0.3, I will introduce the main characters of this investigation—viz. mainstream research on human rationality and the framework of AR. In section 0.4 I will discuss the methodology, originality, and limitations of the present work. The final section (0.5) will also provide a plan of action as well as further details on the structure of the arguments I offer here.

0. 2 A pessimistic view of human rationality

Man is a rational animal—so at least I have been told. Throughout a long life, I have looked diligently for evidence in favour of this statement, but so far I have not had the good fortune to come across it, though I have searched in many countries spread over three continents.
Bertrand Russell

The task of this section is to introduce the target of AR's attack, and to outline the first character of our investigation. Painting with a broad brush, here I will refer to this target as "mainstream bias research" (henceforth, MBR). This tag refers mainly to research carried out in the *heuristics-and-biases* tradition (e.g., Gilovich, Griffin and Kahneman 2002; Nisbett and Ross 1980), which constitutes one of the most successful research projects in the cognitive sciences, as well as the key polemical target motivating the AR theorists' reactions and claims. However, it would be

reductive to associate the claims that I present here with research in such tradition only. Here I wish to present a set of beliefs and contentions that are supposed to capture the key assumptions and premises of much experimental work in social and cognitive psychology. In the remainder of this section I look at the views of scholars working in several different research projects, all concerned with biases affecting human rationality. This should help to show that I am not tilting at windmills here.

With this clarification in place, let me carefully introduce MBR. It is important that we fully understand the concept of bias, since the debate over human rationality revolves around whether and to what extent people are prone to bias. In psychological research, the term bias has been used to refer to systematic deviations from normative standards.[1] But the question that arises at this point is which normative standards should count as the "right" ones. This seems to be a difficult question and choosing the right norms poses a daunting task.

First, it is important to stress that researchers in MBR seem to emphasize the connection between rationality and adaptive and successful behaviour.[2] Notably, most researchers engaged in the empirical study of human rationality subscribe to an instrumental view of rationality, and seem to believe that what determines whether cognition and behaviour are "good" is whether they are conducive to people's goals.[3] These researchers seem to be interested in whether people's behaviour and cognition

[1] For a useful discussion of several different uses of the term "bias" in different disciplines, see, e.g., Hahn and Harris (2014).
[2] This is more controversial in the philosophical literature, and in Chapter 3 we will discuss some attempts to problematize the connection between rational and adaptive behaviour.
[3] The perspective of instrumental rationality is often associated with Hume's work. While exegesis goes beyond the scope of this dissertation, it is useful to highlight that it is debatable to what extent such attribution is accurate (cf. Broome 1999; Sugden 2006).

lead to the achievement of goals and desired outcomes. For instance, Herbert Simon offers an explicit statement of this view when he points out that 'reason is wholly instrumental. It cannot tell us where to go; at best it can tell us how to get there. It is a gun for hire that can be employed in the service of any goals we have, good or bad' (1983, 7–8).[4] However, as David Over points out, the commitment to instrumental rationality is almost universally held in empirical research on human rationality; he tells us that 'almost all discussions of rationality in cognitive psychology and cognitive science presuppose this instrumentalist understanding of rationality' (2004, 5). Unsurprisingly, even contemporary researchers working in the tradition of research in the *heuristics-and-biases* project have offered clear expressions of these commitments. For instance, cognitive psychologists Keith Stanovich and Richard West claim that 'adaptive decision making is the quintessence of rationality' and that 'to think rationally means taking the appropriate action given one's goals and beliefs' (2014, 81). Importantly, a distinction is often made between instrumental rationality (actions that maximize the probability of success) and epistemic rationality (belief formation that is accurate and truth tracking), and the idea that behavior and cognition are successful is typically taken to refer to practical success (achieving one's desires), excluding cognitive aims. While the relationship between instrumental and epistemic rationality is controversial (e.g., Kelly 2003), it is important to stress that, in empirical debates on human rationality, epistemic rationality is often seen as a species of instrumental rationality, namely instrumental rationality in the service of one's cognitive and epistemic goals.

[4] Notably, scholars in different traditions (both MBR and AR theorists) have presented Simon's work as a main source of inspiration.

Second, besides this general commitment to instrumental rationality and such interest in adaptive behaviour and cognition, MBR researchers have generally declared that rational behaviour has to be assessed against norms of first-order logic, probability theory, and rational decision theory. Given the popularity of this view, such norms are often taken to constitute what is referred to as the "standard picture of rationality" (Stein 1996, 4; henceforth SPR). Notably, Edward Stein writes that, according to SPR:

> To be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory and so forth. If the standard picture of reasoning [rationality] is right, principles of reasoning that are based on such rules are normative principles of reasoning, namely they are the principles we ought to reason in accordance with. (1996, 4)

It is worth noting, however, that on some occasions other labels have been used when referring to this normative perspective on rational behaviour and cognition. For instance, Chase, Hertwig and Gigerenzer use the expression "classical view" and write that this view 'equates rationality with adherence to the laws of probability theory and logic [and] has driven much research on inference' (1998, 206). Here I will, at least provisionally, start by referring to Stein's (1996) SPR, since this label has become very popular and is still widely used.[5] I should stress that, for the purpose of this chapter, the characterization of SPR offered by Stein seems to fit the descriptions of methodology and commitments that researchers in the field of judgement and decision-making typically offer. Take, for example, the view of Jonathan Baron, who wrote that 'the major standards come from probability theory,

[5] While Stein refers to this normative perspective as the "standard picture of rationality" (1996, 4) and Chase, Hertwig and Gigerenzer (1998) use the expression '"classical rationality", Evans and Over (1996) seem to characterize this view in terms of "impersonal rationality", Chater and Oaksford (2000, 99) as "formal rationality" and Kacelnik as "axiomatic rationality" (2006).

utility theory, and statistics. These are mathematical theories or "models" that allow us to evaluate a judgment. They are called normative because they are norms' (2004, 19). Moreover, consider the words of Amos Tversky and Daniel Kahneman, pioneers in the *heuristics-and-bias* tradition, who describe their project as relying on the normative rules of 'the modern theory of decision making under risk' (1986, S252), encompassing transitivity of preferences, dominance, invariance, and cancellation. Finally, it seems that Nisbett and Ross had such picture in mind when writing that they 'follow the conventional practice by using the term "normative" to describe the use of a rule when there is a consensus among formal scientists that the rule is appropriate for the particular problem' (1980, 13).

Unsurprisingly, in light of these statements, cognitive errors and biases are explicitly presented in the literature as violations of the norms of SPR. Representative of this trend is the characterization of research on bias offered by Wilke and Mata:

> First, participants were presented with a reasoning problem to which corresponded a normative answer from probability theory or statistics. Next, participants' responses were compared with the solution entailed by these norms, and the systematic deviations (biases) found between the responses and the normative solutions were listed. (2012, 53)

Notably, also philosophers commenting on the "rationality debate" seem to accept this characterization of bias research. For instance, Samuels, Stich and Bishop, when discussing research in the *heuristics-and-biases* tradition, write that these researchers 'appear to be in the business of evaluating the intuitive judgements that subjects make against the standard picture of rationality' (2002, 247). Moreover, Sturm, in a

more general discussion of MBR, writes that 'a massive amount of the psychological literature from the last decades has applied the following procedure: pick a particular norm […] and see how many experimental subjects apply it correctly' (2012, 68).

Third, according to a popular view, following the rules of SPR is tantamount to pursuing adaptive behaviour and cognition, and violating such norms leads to unsuccessful behaviour and cognition.[6] To appreciate this idea, consider how the norms of SPR are often justified. In particular, consider how the importance of the transitivity axiom—perhaps the core pillar of the "standard picture of rationality"—is typically explained. If one satisfies the transitivity axiom, one cannot become a "money pump". Suppose a person prefers option A to B, B to C, and C to A, thus violating transitivity. We can provide this person with option A and then ask, 'Would you pay me a very small amount if I were to replace A with your preferred item C?' If the person agrees, one can then make the same offer concerning the replacement of option C with B, then B with A, then A with C, *ad infinitum*. In this way, the argument goes, a person with intransitive judgment thus becomes a "money pump" who can be exploited by a series of offers. Notably, researchers in the field of judgment and decision-making have often appealed to pragmatic justifications of those norms to defend their reliance on traditional normative requirements.[7] For

[6] It should be noted, however, that some scholars adopting SPR were only motivated by a desire to offer accurate descriptions of human judgment and decision-making and to give insight into underlying mechanisms and processes.

[7] This does not mean that pragmatic justifications of such norms are the only justifications available (cf. Hansson and Grüne Yanoff 2006). In fact, it is worth noting that, especially in other bodies of literature, such as in philosophical treatments of decision theory, these justifications are less popular (cf. Grüne Yanoff 2012). Specifically, it is often argued that it would be misguided to assume that pragmatic considerations provide justifications. Instead, examples such as the Money Pump elicit intuitions, and normative judgments arise directly through human intuition, guided by reflection, and these intuitions, rather than pragmatic considerations, constitute the grounds for normative judgments.

instance, this seems to capture what Amos Tversky and Daniel Kahneman have in mind when they write that 'a common argument for transitivity is that cyclic preferences can support a "money pump", in which the intransitive person is induced to pay for a series of exchanges that returns to the initial option' (1986, S253). Moreover, Keith Stanovich, a cognitive psychologist who has been developing key themes of research in the *heuristics-and-biases* tradition, seems to have a similar idea in mind when he points out that:

> Precisely the reason why people should want to follow the axioms of utility theory (transitivity, etc.) as normative models is that failure to follow them means that a person is not maximizing utility. They should want to avoid becoming a money pump. (2011, 269)

The general idea behind these justifications is that only if—and as long as—following the norms of SPR is conducive to successful behaviour and cognition, do these norms have normative force. This is clearly stated by Jonathan Baron, who points out that:

> If it should turn out that following the rules of logic leads to eternal happiness, then it is "rational thinking" to follow the rules of logic (assuming that we all want eternal happiness). If it should turn out, on the other hand, that carefully violating the laws of logic at every turn leads to eternal happiness, then it is these violations that should be called rational. (2000, 53)

Fourth, researchers in MBR tend towards the view that people frequently and

---

However, it is also the case that justifications of norms of rationality that rest on people's intuitions only have been criticized recently (e.g., Baron 2001; Kahneman 1981). For instance, Weinberg, Stich and Nichols (2001) criticize such approach to the justification of normative principles of reasoning, although the focus in their paper is mainly on some alleged shortcomings of research in traditional analytic epistemology. Specifically, these scholars take issue with what they dub "intuition driven romanticism", viz. the attempt to derive normative claims from epistemic intuitions. They claim that 'perhaps the most familiar examples of intuition-driven romanticism are various versions of the reflective equilibrium strategy' (433), where reflective equilibrium is sometimes presented as a method for the justification of normative principles (cf. Goodman 1965).

systematically violate the norms of SPR. But this point needs to be further clarified. It is true that, in some seminal empirical explorations of human rationality in the field of judgement and decision-making, researchers licensed only moderate verdicts about human irrationality. For instance, research in the 1960s examined the revision of beliefs in light of new evidence. This line of research typically used "bookbags" and "pokerchips", that is, bags containing varying compositions of coloured chips (e.g., 60% red and 40% blue in one bag, 40% red and 60% blue in the other). Participants saw samples being drawn from one of these bags and indicated their new, revised, degree of belief in the composition of chips in that bag (e.g., that the bag had predominantly blue chips). This line of research allowed researchers to careful assess the extent to which participants' belief revision matched the prescriptions of Bayes' rule as a norm for updating beliefs (e.g., Peterson and Uleha 1964; Philips and Edwards 1966; see Peterson and Beach 1967, and Slovic and Lichtenstein 1971 for reviews). The main result of such research was that people responded in qualitatively appropriate ways to evidence, but—quantitatively—did not revise their beliefs as frequently as the normative prescription of Bayes' rule demands. These systematic deviations from optimal responses did not, however, result in researchers forming negative conclusions about human rationality. In fact, the conclusion was that probability theory, taken to provide optimal models for making inferences under conditions of uncertainty, offers 'a good first approximation for a psychological theory of inference' (Peterson and Beach 1967, 42).

Things quickly changed, however, with the development of the *heuristics-and-biases* approach to judgment and decision-making, pioneered by Kahneman and

Tversky. A major element of this approach was the idea that people's cognition is characterized by the use of heuristics or rules of thumb. But if providing a characterization of the notion of bias is complicated, offering an explication of the notion of heuristic can prove to be more complicated still. The term "heuristic" is of Greek origin and means 'serving to find out or discover'; but the original meaning of the term does not seem to be what Kahneman and colleagues have in mind. Kahneman and Frederick (2002) propose that heuristics are shortcuts that people use in making judgements and decisions.[8] Whatever the original intention, this approach has led to the view that heuristics and biases are inseparable twins, to the point that coming up with descriptive models of heuristics has been taken as tantamount to explaining bad reasoning. For instance, in one of the most celebrated studies in the *heuristics-and-biases* tradition inaugurated by Tversky and Kahneman, subjects were presented with a description of some fictional person:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Subjects were then asked to rank the following statements from most to least probable:

> (a) Linda is a teacher in elementary school
> (b) Linda works in a bookstore and takes yoga classes
> (c) Linda is active in the feminist movement
> (d) Linda is a psychiatric social worker

[8] But see Chow (forthcoming) for an interesting critical discussion of different uses of the term "heuristic".

(e) Linda is a member of the League of Women Voters
(f) Linda is a bank teller
(g) Linda is an insurance sales person
(h) Linda is a bank teller and is active in the feminist movement

Strikingly, almost 90% of "naïve" subjects rated (h) more probable than (f), and almost the same pattern has been found among "sophisticated" subjects. This means that a major portion of the subjects rated the probability of the conjunction (a & b) as higher than that of the single event (a), despite the fact that this violates the *conjunction rule* in probability theory. Most people accept that P (a & b) > P (a), whereas probability theory states that P (a & b) ≤ P (a) and P (a & b) ≤ P (b): Linda simply cannot be a feminist bank teller unless she is a bank teller. People's behaviour has been explained by appealing to the *representativeness heuristic*, which refers to making a judgment on the basis of 'the degree to which it is (i) similar in essential properties to its parent population and (ii) reflects the salient features of the process by which it is generated' (Kahneman and Tversky 1972, 431).

Research in this tradition has documented a huge number of pitfalls that seem to prevent our cognition from being optimal. Another striking case in which people's cognition seems to be characterized by errors and biases comes from research on so-called *framing effects*, which occur when alternative frames of essentially the same decision problem lead to predictably different choices. Let us focus here on the Asian disease problem, analysed by Tversky and Kahneman (1981). Experimenters asked test subjects to make two choices with respect to a scenario in which the outbreak of a disease threatens the lives of 600 people.

In the first choice subjects were asked which of the two options would be preferable:

> a) 200 people are saved
> b) There is 1/3 probability that 600 people will be saved and 2/3 probability that nobody will be saved

You may think that the value of saving 200 lives is not equivalent to 1/3 of the value of saving 600 lives. However, the choice can also be framed the following way:

> c) 400 lives are lost
> d) There is a 1/3 probability that no lives are lost and 2/3 probability that 600 lives are lost

The only difference between a/c and b/d is in how the choices are framed: the first choice presents the outcomes in a positive light, whereas the second presents them negatively. However, participants in the study displayed inconsistent patterns of preference: 72% preferred a) to b), while 78% preferred d) to c). There seems to be something going wrong with people's preferences, as the only difference between a/c and b/d is in how the outcomes are framed. A requirement of SPR is therefore being breached. More precisely, the researchers in this study refer to a violation of the *principle of invariance*, which they characterize in the following terms:

> An essential condition for a theory of choice that claims normative status is the principle of invariance: different representations of the same choice problem should yield the same preference. That is, the preference between options should be independent of their description. Two characterizations that the decision-maker, on reflection, would view as alternative descriptions of the same problem should lead to the same choice—even without the benefit of such reflection. (Tversky and Kahneman 1986, 253)

These are two better-known examples of studies seemingly showing that people fail to reason according to SPR. But they are by no means the only studies with results along these lines. For instance, people also commit the *gambler's fallacy*, believing that the probability of some event (e.g., that a coin will come up heads next time it is tossed) is somehow influenced by a certain pattern of previous occurrences (its having come up tails the previous five tosses). So, too, there is *base-rate fallacy*, a clear violation of Bayesian probability calculus: asked to estimate the probability that a given patient has some disease, given a positive result on a test with such-and-such an incidence of false positives, people ignore the information they are explicitly presented with, in this case about the prevalence of the disease in the general population (Casscells, Schoenberger and Graboys 1978). More examples could be offered. But enough has been said to make it at least understandable why, according to several scholars, people are prone to systematic and widespread bias.

Fifth, given the allegedly strong link between adaptive behaviour and cognition and SPR, it should come as no surprise that biases have been taken to be responsible for key cases of unsuccessful and maladaptive behaviour. For instance, biases have been linked to unsuccessful performance in a number of domains, such as human resource decisions (e.g., Highhouse 2008) or health-related decisions (e.g., Reyna and Brainerd 2008). Rather explicitly, Stanovich writes that 'empirical and theoretical work in behavioural finance, much of it using normative approaches, now constantly appear in media articles attempting to explain aspects of the 2008/2009 financial crisis' (2011a, 268-269). Also Milkman, Chugh and Bazerman (2009)

express the view that biases are conducive to maladaptive behaviour. According to these authors:

> Errors induced by biases in judgment lead decision makers to undersave for retirement, engage in needless conflict, marry the wrong partners, accept the wrong jobs, and wrongly invade countries. Given the massive costs that can result from suboptimal decision-making, it is critical for our field to focus increased effort on improving our knowledge about strategies that can lead to better decisions. (379)

But similar claims about the costs of biased thinking and decision-making are not hard to find. For instance, Johnson and Levin write that:

> In an ideal world, people would tackle major crises such as global climate change as rational actors, weighing the costs, benefits and probabilities of success of alternative policies accurately and impartially. Unfortunately, human brains are far from accurate and impartial. Mounting research in experimental psychology reveals that we are all subject to systematic biases in judgment and decision-making. […] In today's world of technological sophistication, industrial power and mass societies, psychological biases can lead to disasters on an unprecedented scale. (2009, 1593)

At times, researchers in MBR have also tried to link specific biases to more specific instances of maladaptive behaviour. For instance, Johnson and Fowler claim that 'overconfidence has been blamed throughout history for high-profile disasters such as the First World War, the Vietnam war, the war in Iraq, the 2008 financial crisis and the ill preparedness for environmental phenomena such as Katrina and climate change' (2011, 317). Interestingly, the fact that these violations of SPR seem to lead to poor outcomes explains why concrete measures have recently been taken to improve people's strategies and behaviour: noting that widespread and systematic errors are conducive to maladaptive behaviour, researchers have claimed that 'the time has come to move the study of biases in judgment and decision making beyond

description and toward the development of improvement strategies' (Milkman et al. 2009, 379). In particular, the claim that people are prone to costly biases has been used to justify public interventions to de-bias them.[9] These interventions—called "nudges"—have been increasingly important in public policy, especially in the UK, with initiatives such as the "Behavioural Insights Team"—a team of behavioural economists who advocate non-coercive policies (Cabinet Office Behavioural Insight Team 2010; Dolan et al. 2011).[10]

In light of these considerations, we can now understand why psychologists such as Nisbett and Borgida write that the prospects for human rationality might be rather 'bleak' (1975, 935). Humans, as such theorists see it, are irrational, since people are prone to systematic and widespread bias. Notably, 'bleak' assessments of human rationality are now widespread, and the scientific impact of research on human bias was formally recognized in 2002 when the Nobel Memorial Prize in Economics was awarded to cognitive psychologist Daniel Kahneman.

[9] For a comprehensive list of "nudges", see: www.stir.ac.uk/media/schools/management/documents/economics/Nudge%20Database%201.2.pdf.

[10] It is worth noting, however, that, whilst nudges are typically defended by pointing out that errors and biases are widespread and costly, one may accept that people are prone to costly errors and still argue against the use of nudges. For instance, one might appeal to one particular line of criticism, according to which these policies are not libertarian (cf. Hausman and Welch 2010, Grune-Yanoff 2012).

0. 3 The challenge from adaptive rationality

People are not logical, they are psych*ological. Anonymous*

A foolish consistency is the hobgoblin of little minds, adored by little statesmen and philosophers and divines. With consistency, a great soul has simply nothing to do.
Ralph Waldo Emerson

This perspective on human rationality adopted in MBR, which I outlined above, has been dominant for many years. This does not mean, however, that it has not attracted criticism. While providing a history of challenges to bleak assessments of human rationality goes beyond the scope of the present work, it is worth mentioning that such criticisms have been rather fierce. For instance, Lopes (1991) claimed that 'the view that people are irrational is real in the sense that people hold it to be true. But the reality is mostly in the rhetoric' (1991, 80).[11] Only recently, however, have a number of researchers started to articulate a new and alternative framework for the study of rational behaviour and cognition: a number of cognitive psychologists and evolutionary behavioural scientists have argued for a paradigm shift in the study of human rationality.

Whilst in several corners of research on judgement and decision-making it is business as usual, one perspective that is becoming more and more popular takes organisms to be matched to the demands of their environments. In much the same

[11] The reader can find some useful discussions of classical arguments against MBR in Stich (1990), Stein (1996), Evans and Over (1996), Hastie and Dawes (2001), and Samuels, Stich and Bishop (2002).

way that biological devices are matched to their niches, decision-making mechanisms are also matched to particular kinds of task. As in the case of biological fit, the suitability of these cognitive mechanisms is predicated on the structure of their environment. For instance, in a recent paper, evolutionary psychologists Hugo Mercier and Dan Sperber wrote that 'human reasoning is not a profoundly general mechanism; it is a remarkably efficient specialized device adapted to a certain kind of information at which it excels' (2011, 72). Against MBR, it is now frequently contended that 'the frequent labelling of behaviours as irrational, anomalies or biases assumes a particular perspective on rational norms' (Stevens 2008, 295) and that 'many common biases and heuristics reflect a deeper adaptive rationality' (Kenrick et al. 2012, 23). Unfortunately, whilst the idea of "adaptive rationality" is now becoming quite popular in the literature as a new perspective on rational behaviour and cognition, these attacks on MBR are rarely spelled out in detail.

Here I will focus on the arguments, claims, and challenge put forth by Gerd Gigerenzer, Ralph Hertwig, and their co-workers at the Centre for Adaptive Behaviour and Cognition (ABC) and at the Centre for Adaptive Rationality (ARC), since they have articulated the perspective in most detail. I will henceforth refer to this relatively close-knit group of researchers when using the term AR.[12] At the same time, it is important to note that the discussion here might have broader implications

[12] It is important to note that what I call AR is often referred to as "ecological rationality" (e.g. Todd and Gigerenzer 2012; Rich 2014). However, I am not alone in preferring the label "AR" to "ecological rationality" (cf. Schurz 2014; Lin 2014, 667). Two reasons for this preference seem to be the following. First, the label "ecological rationality" is often used to refer to the study of the match between decision mechanisms and the environment, which is not clearly a normative project. Second, other researchers outside AR (e.g., Smith 2003) have appealed to the notion of "ecological rationality" to refer to normative views that differ from those I associate here with AR theorists.

for the "rationality debate", since a growing number of researchers seem to share such an optimistic perspective on human rationality (e.g., Cosmides and Tooby 1994; Mercier and Sperber 2011).

Notably, AR researchers stress that 'we need not worry about human rationality' (Gigerenzer 1999, 280). The AR project has been often presented as attempting to show 'how people are able to achieve intelligence in the real world' (Todd and Gigerenzer 2012, 3), and provocative book titles such *Adaptive thinking: Rationality in the Real World* (Gigerenzer 1999) or *Ecological Rationality: Intelligence in the Real World* (Todd and Gigerenzer 2012) illustrate quite clearly the general aims of their project. This optimistic view is motivated, in part, by the contention that behaviour and cognition have been assessed in rather abstract contexts, which are not representative of the real world. More importantly, however, researchers in MBR are charged with assessing behaviour against the wrong normative standards. Here I will try to clarify the nature of the AR theorists' concerns with the SPR and to localize the conflict between MBR and AR.

First of all, it is important to stress that AR scholars highlight the importance of adaptive behaviour and cognition and embrace an instrumental view of human rationality.[13] For instance, it is claimed that:

[13] Note, however, that this is a contentious point. Stanovich and West (2003) have charged evolutionary psychologists and AR theorists with departing from instrumental rationality and focusing only on the question of whether behaviour is evolutionarily adaptive. Whilst this seems to be true for several evolutionary psychologists, the objection does not seem to apply to AR theorists. And this is

> The rationality naturalized by our psychological research program picks out one specific sense of rationality, albeit an important one. We start from a kind of means-ends rationality—the "default" notion of rationality. (Gigerenzer and Sturm 2012, 245)[14]

This is clearly an important commitment that AR seems to share with MBR. It is also interesting to note that this is a non-trivial assumption. In fact, in philosophical debates in particular, it has been argued that this sort of rationality is overly limited. For many, the key question a theory of rationality has to answer is what ends we ought to pursue and which goals it is rational to have. To give an example, according to instrumental rationality a person who wants to drink a can of paint and chooses the best means towards this— opening the can in the appropriate way— is acting rationally. However, to many there is nothing rational about someone who efficiently goes about drinking paint (e.g., Richardson 1994; Korsgaard 1997). The general point is that genuine rationality requires the adoption of rational ends, such that the choice of means is not sufficient. What is worse, according to critics of instrumental rationality, is that the most interesting aspect of rationality is simply missing from theories of instrumental rationality. Here I will not seek to defend the commitment to instrumental rationality. Rather, I will simply note that this commitment is not what marks the contrast between AR and MBR: key figures engaged in psychological

---

one reason for my focusing on the challenge mounted by AR theorists here. Specifically, while it is true that AR theorists have often appealed to evolutionary considerations, they have done so mainly to account for the performance of people's heuristics, and not to find new benchmarks of rationality. In fact, AR theorists have explicitly distanced themselves from this position, stressing that 'the study of ecological rationality […] should not be confused with the biological concept of adaptation' (Gigerenzer and Gaissmaier 2011, 458) and explicitly endorsing the perspective of instrumental rationality (cf. Gigerenzer and Sturm 2012). In the following chapters, I seek to show that AR theorists do not merely pay lip service to instrumental rationality, but seem to be genuinely interested in people's ability to achieve their goals (not only evolutionary goals).

[14] Notably, AR theorists make clear that they do not only care about the fulfilment of desires, but also about the achievement of epistemic goals (cf. Gigerenzer and Sturm, 2012, 254).

research on thinking and decision-making have traditionally been concerned with the achievement of goals.

Instead, what sets the AR project apart from MBR is the idea that following the norms of SPR is functional to the pursuit of adaptive behaviour and cognition, and that violating the tenets of SPR leads to maladaptive behaviour. Specifically, AR theorists reject the key claim of MBR, namely that the norms of SPR should be used as benchmarks of rational behaviour and cognition. In fact, AR theorists stress that:

> Looking at the relation of heuristics to environments is often normatively more useful than evaluating reasoning and decision-making according to the standard norms of probability or decision-theory. (Gigerenzer and Sturm 2012, 254–255)

At other times, AR theorists present their perspective as more explicitly being 'in stark contrast to classical definitions of rationality' (Rieskamp and Reimer 2007, 273). This does not mean that, according to AR theorists, following the norms of SPR is wrong in all contexts. But in several contexts it is: what this means is that behaviour departing from the norms of SPR is successful in a number of domains, and that as a consequence these norms should not be used as benchmarks of rational behaviour and cognition.[15]

[15] Besides the questions of how to assess behaviour alongside the question of whether human behaviour and cognition are adaptive, there is the question of how this adaptivity could be achieved. With regard to this last question, AR theorists have articulated a framework known as "adaptive toolbox", according to which decision-makers will select among a set of strategies, and different rules describe the actual information integration mechanisms. While in Chapter 3 I will present and critically discuss the "adaptive toolbox" framework, there are alternative hypotheses and a huge literature on the plausibility of single-mechanism and multi-strategy accounts that I will not be able to discuss in this work. The reader can refer to Glöckner et al. (2014) for an up-to-date assessment of the framework of the "adaptive toolbox" and a comparison with alternative hypotheses.

0.4 The methodology and scope of the project

This thesis provides one of the first philosophically-informed appraisals of the prospects of the AR project.[16] But this claim needs to be qualified. In fact there are several ways in which one might look at this debate from a philosophical perspective. After all, rationality is central to our self-conception, and claims about human rationality figure prominently in philosophical theories concerned with human action, such as ethics, where theories are often constructed with a specific kind of agent in mind, namely the rational agent.[17]

Here I will not discuss all the philosophical issues underlying the empirical debate on human rationality or raised by it. Instead, I will look at the key claims and arguments made by MBR and AR theorists, and try to reconstruct them as accurately as possible, assessing them against the empirical evidence already available. In so doing, I will call on evidence from scientific disciplines such as social and cognitive psychology, economics, and evolutionary biology. Overall, the present work will be an instance of what is often referred to as "empirical philosophy" (Prinz 2008), where this refers to using empirical facts in one's philosophical theorizing. More specifically, what I will do here is to attempt to contribute to scientific theorizing by providing some novel hypotheses and original speculations, by synthesizing swath of empirical and theoretical works, and by suggesting empirical research.

[16] For other philosophically-informed discussions of some aspects of AR, see, e.g., Arnau et al. (2014), Bortolotti (2011; 2014), Hurley (2005) and Samuels et al. (2001).
[17] For example, an interesting discussion of the implications that empirical research on human rationality can have for ethics can be found in Brännmark and Sahlin (2010).

But more has to be said in order to clarify the nature of my approach. In particular, it is important to highlight that my contribution to the "rationality debate" departs in important ways from other philosophical treatments. First, too often philosophers discussing the AR project have either failed to focus on key normative issues arising from the AR challenge or have mischaracterized the AR project. More precisely, several philosophers have focused on descriptive or methodological aspects of the challenge and left key normative considerations aside (e.g., Sterelny 2003; Schulz 2011a).[18] On some other occasions, a number of philosophers have explicitly claimed that the debate between AR theorists and researchers in MBR rests on minor nuances rather than genuine disagreements. A very clear example of this trend can be found in Samuels, Stich and Bishop's article *Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear* (2002), which has been well received in the literature (cf. Burns 2004, 326; Grüne-Yanoff 2007, 554; Rysiew 2008, 1172; Sinnott Armstrong et al. 2010, 249). According to these authors, the clashes between these two different perspectives and interpretations are mainly due to a failure to distinguish between the 'core claims' and the 'rhetorical flourishes' of the competing research programs. As soon as this distinction is recognized, it is possible to realize that there is no real disagreement in the debate. In particular, according to these authors, there are no genuine disagreements at the normative level between the two parts in this dispute, as researchers in MBR as well as AR theorists 'typically presuppose what Edward Stein has called the standard picture of rationality' (2002, 253). And while scholars in these different research

[18] It is worth mentioning, however, that other authors have focused on a rather narrow set of issues arising from the normative challenge mounted by AR (Vranas 2000).

programs have made rather different claims about human rationality, the overly optimistic or pessimistic claims about human rationality expressed by AR theorists and MBR researchers respectively should be classed as mere 'rhetorical flourishes'. Following this interpretation offered by Samuels et al. (2002), Rysiew claims that 'the divisive rhetoric of the rationality wars masks what is in fact a deeper significant agreement' (2008, 1172).[19] The approach adopted in this thesis is radically different. I take the challenge mounted by AR theorists to be the most radical and sophisticated challenge to MBR currently available, and dedicate the key chapters of this work to an assessment of AR theorists' normative claims, as AR theorists' normative perspective departs from MBR researchers' commitment to SPR.[20]

But there is a second popular trend in philosophy from which I would like to distance my work. Specifically, philosophers interested in rational behavior and cognition have typically focused on *a priori* arguments.[21] Consider, for instance, the so-called "interpretationist argument" typically attributed to Davidson (1984) and Dennett (1987). According to this argument, considering people as rational is a

[19] Yet this tendency to reduce the "rationality debate" to marginal or superficial disagreements has also extended beyond the philosophical literature. For instance, cognitive psychologist Burns writes that the 'differences between the *heuristics-and-biases* approach and the adaptive approach are minimal' (2004, 49).

[20] Interestingly, AR theorists themselves often lament that their perspective has been misrepresented. For instance, Gigerenzer writes that 'the story is told that there are two personalities among psychologists, optimists and pessimists, who see the glass as half full or half empty, respectively. According to this legend, people like Funder, Krueger, and myself are just kinder and more generous, whereas the pessimists enjoy a darker view of human nature. This story misses what the debate about human irrationality is about. It is not about how much rationality is in the glass, but what good judgment is in the first place. It is about the kinds of questions asked, not just the answers found' (2004, 26).

[21] It is worth noting, however, that things have been gradually changing, as a growing number of philosophers are now taking empirical research on judgement and decision-making very seriously. For instance, Lisa Bortolotti writes that 'when philosophers address traditional questions such as: 'Are humans rational?' […], they can benefit from paying attention to research in cognitive science' (Bortolotti 2011, 297).

precondition of crediting them with intentional states and engaging in reasoning. For another case, consider the argument offered by Cohen (1981), who stresses that human irrationality cannot be experimentally demonstrated. Specifically, according to his argument, our views regarding rationality should be reached through a process of adjustment between our intuitions about the rightness of particular inferences and our assessment of proposed general principles of rationality, with the goal being to obtain a systematization of the latter. In this process, there is no higher court of appeal than humans' intuitions about various principles and cases. In Cohen's own words 'a normative theory of rationality must be based ultimately on the data of human intuition' (321). The outcome, according to Cohen, is that the reasoning errors people seem to make *must* be considered as mere performance errors, which do not impugn the reasoning *competence* that all normal humans possess (321).[22]

Such an emphasis on *a priori* arguments seems to mark, at least to some degree, a difference in attitudes exhibited by philosophers and psychologists interested in the study of human rationality. But this focus on *a priori* rather than empirical issues has also resulted, in turn, in another interesting difference between the two approaches. By focusing on *a priori* arguments, a number of philosophers have taken MBR to provide little insight into people's alleged irrationality. Notably, as Botterill and

[22] These arguments have been extremely influential in the philosophical literature, but—perhaps unsurprisingly—experimental researchers have not taken them equally seriously. For instance, Shafir and Leboeuf write that 'the status of the rationality assumption is ultimately an empirical question (but see Cohen 1981, Dennett 1987). Consequently, the field of experimental psychology has been at the forefront of the modern rationality debate' (2002, 492). For discussions of the abovementioned philosophical arguments, see Stich (1990) and Stein (1996). Moreover, interesting criticisms of Cohen's (1981) argument can be found in the commentaries on his BBS paper. Bortolotti (2005) offers some criticism of the "interpretationist argument". In addition, for an empirically-informed discussion of the plausibility of Dennett's and Davidson's views on people's attribution of mental states, see, e.g., Nichols and Stich (2003) and Goldman and Mason (2007).

Carruthers (1999, 105) have pointed out, the disciplinary division between psychologists and philosophers marks two different attitudes because psychologists seem to find only grounds for pessimism about human reasoning and decision-making, whereas philosophers appear far more confident about human rationality. Botterill and Carruthers' insight is, at least to some extent, correct, since the irrationality thesis is more widely held among psychologists than among philosophers.[23]

Contrary to this trend, in this work I assume that there are a number of empirical findings that are worthy of attention and likely to have a bearing on the plausibility of attributions of (ir)rationality. Here I focus on evidence concerning the relationship between following the norms of SPR and achieving successful and adaptive behavior, as I take these findings to be the most relevant and important. But there are other empirical findings besides those I discuss here that might have a bearing on the "rationality debate". For instance, AR theorists have recently given other evidence in support of their case, stressing, in particular, the puzzling fact that mental abnormalities seem to be associated with conformity to SPR (Hertwig and Volz 2013). More precisely, individuals with mental illnesses or damage to specific brain regions are more likely than healthy individuals to adhere to SPR. For example, patients with damage to the ventromedial prefrontal cortex (VMPFC) are more coherent in their preferences in a consumer choice context (Koenigs and Tranel

[23] Yet, as one might imagine at this point, a number of psychologists have also started to embrace rather optimistic views on human rationality. In addition to the literature already mentioned, Osman (2014) has recently articulated a perspective on how we might guide behavior change that is more optimistic than those suggested by recent books like *Nudge* (Thaler and Sunstein 2008) and *Predictably Irrational* (Ariely 2009).

2008). Likewise, they do not seem to fall prey to the *correspondence bias*, which means they are less likely to assume that outcomes are caused by dispositional factors, e.g., a person's constitution or personality, even when the actual cause is due to situational factors (Koscik and Tranel 2013). Moreover, there are other empirical challenges and puzzles for MBR that do not come from the AR project. For instance, some scholars have pointed out that there appears to be a striking (and puzzling) dissociation between human perceptuo-motor and cognitive decision-making performance. Specifically, whilst human high-level cognitive decisions appear to be sub-optimal, paradoxically, perceptuo-motor decisions appear to be nearly optimal (e.g., Jarvstad et al. 2014; Trommershäuser, Landy and Maloney 2006; 2008). As Trommershäuser, Landy and Maloney point out, for example, 'in marked contrast to the grossly sub-optimal performance of human subjects in traditional economic decision-making experiments, our subjects' performance was often indistinguishable from optimal' (2006, 987). Finally, a number of researchers are now challenging claims made by researchers in MBR by appealing to considerations from cognitive modeling. In particular, a number of theorists have tried to model human cognition according to quantum probability theory,[24] arguing that the latter predicts many of the standard anomalies discussed by researchers in MBR and raising some fundamental questions about the value of SPR (Pothos and Busemeyer 2013; 2014; Wendt 2015, chap. 8).

[24] Quantum probability is a formal theory of probability and an alternative to classical probability theory. Specifically, quantum probability refers to the mathematics for assigning probabilities to events from quantum mechanics, without the physics.

Whilst findings from other research programs might turn out to be interesting, it seems that only in AR do we already have a fully-fledged alternative picture of rationality and a project that seeks to replace both the methodology and the traditional normative standards of MBR. This is why, despite the existence of several empirical challenges to MBR, this enquiry focuses only on that from AR. Note, for instance, that while authors such as Mercier and Sperber (2011), Johnson and Fowler (2009) and Johnson et al. (2013) have mainly focused on the *confirmation bias* and on the *overconfidence bias* respectively, the attack mounted by AR researchers is supposed to apply to research on biases more generally. In addition, while other projects are trying to open up new fronts in the "rationality debate", suggesting, as we have seen above, that human cognitive processes might obey quantum rather than classic (Bayesian) probability theory (Pothos and Busemeyer 2013), those scholars—unlike AR theorists—are generally more reluctant to argue in favor of a wholesale replacement of SPR (e.g., Pothos and Busemeyer 2014).

0.5 Plan for action

> *"Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop."*
>
> Lewis Carroll, Alice in Wonderland

Having provided a description of this dissertation—delineating what it is and is not about—and having presented the questions that motivated this work, the time is now ripe for introducing my argument in more detail. The most original ideas put forth by

this project can be stated as follows. AR theorists have successfully shown that formal principles of rationality cannot and should not be used as universal benchmarks of rationality for the study of adaptive behaviour and cognition, and that researchers should try to assess behaviour against the goals people entertain. However, accepting this claim, which amounts to a conceptual revolution in the study of human judgement and decision-making, does not imply that we should draw optimistic verdicts about human rationality. In fact, there is a great deal of evidence suggesting that people are often unsuccessful at achieving prudential and epistemic goals. I argue that AR theorists have not made clear how, in light of this evidence, optimistic claims about human rationality could be defended. Thus, my thesis provides only a qualified defence of the challenge from AR.

This dissertation, which contains seven chapters, an introduction and a conclusion, is structured as follows.

In Chapter 1, I critically discuss the claim that biases disappear in the "real world". AR theorists have stressed on a number of occasions that many violations of SPR found in the lab are not representative of behaviour in the real world. I challenge their criticism of MBR and show that concerns about the external validity of findings in MBR do not warrant the rejection of pessimistic assessments of human rationality. More specifically, I make a three-pronged attack. First, the evidence about the robustness of the effects discussed by AR theorists is mixed. Second, the contexts that AR theorists class as unrepresentative of the real world are—at least to

some extent— representative of it. Third, many instances of biased behaviour have been studied in the wild, not just in the laboratory. Overall, it seems that the claim that biases disappear in the real world is unsupported: people indeed seem to commit systematic violations of SPR in the real world.

However, the conclusion of Chapter 1 might seem to be at odds with evolutionary considerations. The idea, in brief, is that if we were prone to systematic and widespread biases, we would then fail to navigate the world successfully and thus would not have evolved. Since—arguably— evolution cannot be questioned, this consideration seems to be problematic for MBR. Chapter 2 discusses this evolutionary puzzle and shows that the conclusion of Chapter 1 is not necessarily at odds with evolutionary considerations, and that we can provisionally retain the conclusion of Chapter 1. In particular, I appeal to the fact that natural selection is not the only cause of evolution and, more importantly, that inaccurate reasoning can be evolutionarily adaptive.

However, the fact that cases of inaccurate reasoning can be evolutionarily adaptive raises a set of questions about the value of the norms of SPR traditionally endorsed in MBR. Chapter 3 addresses a number of concerns in this vein, and seeks to show that, at least in a significant number of domains, behaviour that departs from the norms of SPR can be adaptive and successful. This result seems to be problematic for scholars in MBR, since they have advocated that such norms should

be treated as universal benchmarks of rationality for the study of adaptive and instrumentally rational behaviour.

While Chapter 3 reveals that the claims of AR theorists have weight, in Chapter 4 I prompt AR theorists to build their challenge on more solid conceptual grounds. In particular, AR theorists have sought to explicate their normative challenge to MBR by appealing to Hammond's distinction between coherence and correspondence criteria of rationality. But this chapter shows that this distinction does not best explicate the challenge AR theorists are talking about, and that a distinction between rule-based and goal-based rationality should be preferred.

Chapter 5 discusses a possible objection to the account of AR I present in this work. It might seem that I have overlooked a crucial worry shared by AR theorists, viz. that the norms of SPR, or what I call "rule-based rationality", are too demanding, because of our cognitive limitations. Interestingly, commentators on the "rationality debate" have typically appealed to versions of the *ought-implies-can* principle to account for the normative relevance of research on human cognitive limitations. This chapter shows that the framework of AR offers a way of interpreting the normative significance of literature on cognitive limitations without being committed to versions of the *ought-implies-can* principle.

In Chapter 6 I reconstruct and assess a defence of rule-based rationality and, in turn, of MBR. It is currently popular to claim that findings on individual differences in judgement and decision-making challenge the plausibility of the AR project. Here I reconstruct and discuss two arguments based on such research. First, reported heterogeneity in the use of heuristics seems to be at odds with the adaptationist underpinnings of the AR project. Second, the existence of correlations between cognitive ability and susceptibility to bias suggests that rule-based rationality is, after all, normatively adequate. I argue that, as things stand, neither of these arguments is compelling.

While my treatment towards AR has been sympathetic so far, in Chapter 7 I also argue that there are reasons to reject its most ambitious claims. Specifically, even if we agree with AR theorists on the importance of assessing behaviour against the goals that people entertain, their claim that people are generally successful at achieving their goals seems problematic. This chapter challenges their rather optimistic claims about human rationality, and to do so, I begin by noting that, while it seems true that behaviour that violates rule-based rationality can be successful when measured against goal-based rationality, this result holds only for some contexts. More importantly, however, I show that many families of biases reported in MBR seem to be instances of unsuccessful behaviour measured against prudential or epistemic goals, and I argue that AR theorists have not made clear how, in light of this evidence, optimistic claims about human rationality could be defended.

The concluding chapter provides a summary of the main claims and conclusions defended in this work.

Chapter 1: Cognitive biases in the "real world"

As we have seen, scholars in MBR have argued that human beings are prone to systematic and widespread biases. In particular, these researchers often point to evidence showing that people's intuitions about probability deviate dramatically from the dictates of probability theory (e.g., Gilovich et al. 2002). However, while scholars in MBR have claimed on several occasions that human beings are 'a species that is uniformly probability-blind' (Piattelli Palmarini 1991, 35), others have expressed worries about the robustness of such biases. AR theorists have offered empirical evidence to support their claim that biases tend to disappear in the "real world". Here I first introduce the line of argument offered by AR, and then show that this version of the challenge to MBR is not as yet particularly convincing. More precisely, although this work on the part of AR researchers is important, their arguments hardly support the claim that biases disappear in the real world. I articulate a three-pronged reply to their argument. First, biases are more robust than AR theorists suppose. Second, the experimental contexts used in MBR are more representative of the real world than AR theorists suppose. Third, evidence about the existence of biases comes from "field experiments" as well, and not only from laboratory studies, and thus seems to be less vulnerable to AR theorists' objections than AR theorists believe. I will begin in sections 1 and 2 by offering a careful introduction of AR theorists' concerns over the robustness of findings in MBR. Later, in sections 3, 4, and 5, I will then articulate my replies to the AR theorists' argument in detail. I will conclude in section 6. The reader more familiar with

general concerns about the robustness of biases can skip sections 1 and 2 and move directly to section 3.

1. An ecological perspective on thinking and decision-making

AR theorists take issue with the methodology used in MBR. But their criticisms reflect a more general concern they have with large areas of psychological research. Specifically, they seem to be attacking a widely shared picture, according to which psychological processes can be explained and studied largely in isolation from their environment. According to such a view, inputs and outputs of a psychological process are located in the environment, yet the explanation of psychological mechanisms is a story about internal activity. Just as an explanation of the mechanisms of a personal computer is typically given irrespective of the environment in which the computer is located, so the explanation of the mechanisms of human psychology should be given largely irrespective of the environment in which the organism is located.[25] This picture makes psychological science especially apt for laboratory work and analyses that abstract psychological processes from their natural environment.[26]

However, this picture has recently come under attack. AR theorists have explicitly attacked this general tendency to ignore ecological factors in psychological research. They stress that:

[25] For a defence of methodological solipsism, see Fodor (1980).
[26] See Chirimuuta and Gold (2009, Chap. 9).

> Ecological perspectives are still rare in cognitive science aside from a few exceptions, such as the perspectives of Brunswik, Gibson, Shepard, and Anderson and Schooler. Most theories are about mental processes only: neural networks, production rules, Bayesian calculations, or dual-systems notions (Gigerenzer 2008, 22).

Ignoring ecological considerations is highly problematic because the organism and the environment are strongly interacting systems. Whilst the AR theorists' target is not confined to MBR, their concerns apply in particular to research on biases. To articulate their perspective, AR scholars often appeal to Herbert Simon's scissors metaphor, according to which 'human rational behaviour is shaped by a pair of scissors whose two blades are the structure of task environments and the computational capabilities of the actor' (1990, 7). Just as we cannot understand how scissors cut by looking at one blade, we will not be able to understand human reasoning by studying either the individual agent or the environment alone. According to AR theorists, the problem with MBR, therefore, is that its scholars have focused on people's heuristics and on the cognitive limitations of the human mind, without taking seriously the nature of the environments people inhabit.

In articulating such concerns, AR theorists have also looked at Egon Brunswik's work on perception (e.g. 1957),[27] which they explicitly mention as a source of inspiration (Gigerenzer, Hoofrage, Kleinboelting 1991). For Brunswik, the basic problem for psychology was the way in which an organism adapts to its environment. To address this question, Brunswik maintained that psychology must view the organism and the environment as interacting systems that, when considered

[27] For a thorough presentation of Brunswik's work, see Hammond and Stewart (2001).

relationally, have the essential characteristic of a 'coming to terms' (Brunswik 1957). Brunswik was critical of the typical design of experimental psychology. Specifically, his work can be seen as an attack on the external validity of much psychological research of his time (although the label 'external validity' was first used by Campbell only in 1957, and came to define the experimental vocabulary in psychology in the following decades).[28] The design typically used in psychological research was one of betting on internal validity, that is, on the sound demonstration that a causal relationship exists between two or more variables, rather than on external validity, which refers to the generalizability of the causal relationship between the experimental contexts. After all, the logic goes, if internal validity is not guaranteed, no conclusions can be drawn about the effect of independent variables. Brunswik introduced the notion of "representative design" to refer to an experimental design that aims at a veridical representation of the environment in which organisms naturally perform. According to Brunswik, researchers must ensure 'that the habitat of the individual, group or species is represented with all of its variables, and that the specific values of these variables are kept in accordance with the frequencies in which they actually happen to be distributed' (1944, 69). For Brunswik, the key concern was being able to generalize results to the person's natural habitat. This matters rather a lot, on Brunswik's view, since people's cognition can be rather accurate in the natural environment in virtue of the exploitation of regularities in that environment. On the other hand, cognition might look flawed in the laboratory, but only because the controlled laboratory setting destroys these regularities.

[28] It is important to note, however, that as the discussion about this threat has started to disseminate into other social sciences, new terms have been used to define the problem. "External validity", "extrapolation", "parallelism" and "ecological validity" have been used in various fields. In the philosophical literature, however, the terms "external validity" and "extrapolation" are those most frequently used (e.g., Guala 2005; Steel 2008).

Drawing on these Brunswikian themes, AR theorists claim that MBR has mistakenly studied cognition without paying attention to the actual environments where thinking and decision-making occur and, as a consequence, that the experimental findings suggesting human irrationality that are celebrated in MBR cannot be generalized to the real world. According to these theorists, the biases reported in the literature are, at least in significant part, attributable to the use of artificial and unrepresentative settings. AR theorists look at Brunswik's work on perception to explicate their concerns about the external validity of findings on biases:[29] it is vital to assess whether in the experimental context the habitat of the individual is represented in an accurate way. Researchers should thus be careful when making generalizations about people's irrationality, in order to not mistake exceptions due to the use of unnatural contexts for the rule. In the words of AR theorists:

> Just as vision researchers construct situations in which the functioning of the visual system leads to incorrect inferences about the world, researchers in the *heuristics-and-biases* program select problems in which reasoning by cognitive heuristics leads to violation of probability theory. However, the conclusions they draw from such unrepresentative designs often differ sharply from those drawn by researchers of perception. Vision scientists do not conclude from the robustness of the Müller-Lyer illusion, for instance, that people are generally poor at inferring object lengths. However, many advocates of the *heuristics-*

[29] It should be noted, however, that AR theorists have also attacked the internal validity of findings from MBR on some occasions (cf. Gigerenzer 2007). For instance with regards to the *conjunction fallacy*, they sometimes subscribe to Fiedler's concerns that 'the [conjunction] fallacy may represent a verbal misunderstanding of the probability concept. [...] The prevailing statistical interpretation of probability (as relative frequency) does not appear to apply to colloquial language because everyday experience is seldom based on semantic frequency counts. Rather, the usual interpretation of "probability" may come close to such subjective criteria as "believability", "degree of confidence", "imaginability" or "plausibility"' (1988, 123–124). There is a huge literature on these concerns (e.g., Politzer and Noveck 1991; 2004; Schwarz 1994) that I do not discuss in this work, but see, e.g., Moro (2009), for a very clear discussion of some of these arguments and concerns.

> *and-biases* program conclude from the cognitive illusions found in laboratory tasks that human judgment is subject to severe and systematic biases that compromise its general functioning. (Chase, Hertwig and Gigerenzer 1998, 206)

This criticism of MBR is also explicitly voiced in the following passage:

> These phenomena have been described as cognitive illusions, and these demonstrations of irrationality are explained in terms of heuristics on which people, equipped with limited resources, need to rely when making inferences about an uncertain world. [...] However, concerns about whether studies demonstrating irrationality preserved an isomorphism between environmental and experimental properties have given rise to a Brunswikian perspective. (Dhami, Hoffrage and Hertwig 2004, 972)

At this point, it should be quite clear that the AR theorists' project seems to be an attempt to extend Brunswik's approach to the study of rational behaviour cognition. It is worth noting, however, that the considerations offered by AR theorists are also consonant with recent trends in philosophy and cognitive science, although they themselves do not acknowledge this point. For instance, consider that there have recently been important shifts in the philosophy of mind towards a view of cognition as (to cite the current slogan) 'embodied, embedded, enactive, and extended'. Andy Clark, for example, has argued that a proper assessment of human cognitive competence cannot overlook environmental factors, for 'advanced cognition depends crucially on our ability to *dissipate* reasoning: to diffuse achieved knowledge and practical wisdom through complex social structures, and to reduce the loads on individual brains by locating those brains in complex webs of linguistic, social, political and institutional constraints' (1997, 180). On this view, as well as in the view of AR theorists, ignoring environmental factors carries the risk of missing the bigger picture of human cognition. For instance, Clark emphasized the importance of

considering the environment in which cognition takes place by referring to the so-called "007 principle":

> In general, evolved creatures will neither store nor process information in costly ways when they can use the structure of the environment and their operations upon it as a convenient stand-in for the information-processing operations concerned. That is, know only as much as you need to know to get the job done. (1989, 64)

As this quote suggests, there are interesting parallels between the AR theorists' project and recent trends in embedded and extended cognition.[30] In fact, some authors have recently highlighted the existence of these connections (Arnau, Ayala and Sturm 2014). At the same time, while proponents of embedded and externalist accounts of cognition have typically sought empirical support from research on memory (Clark and Chalmers 1998; Sutton et al. 2010) and perception (Wilson 2010), the literature does not typically focus on thinking and decision-making (but see Clark 2001). In this sense, this aspect of the AR project can be seen as complementary, since they focus on higher cognition and, more precisely, on rational judgement and decision-making.

[30] It is important to note, however, that, unlike AR theorists, a number of scholars within the broad framework of situated cognition tend to make quite radical and revolutionary claims about the nature of the cognitive processes that lead to adaptive behaviour. As AR theorists Henry Brighton and Peter Todd point out, 'the approach we advocate here is conservative in comparison with other more radical situated positions. […] Ecologically rational heuristics are uniformly described in terms of symbolic process models operating on representations. These processes draw on the classical notions of search, satisficing, and decision rules. In contrast to more radical positions, the concept of ecological rationality is agnostic with respect to, for example, issues of antirepresentationalism (Slezak 1999; Varela, Thompson and Rosch 1991), dynamic systems theory (van Gelder 1995), or more philosophical rethinkings of the nature of cognition (Winograd and Flores 1986)' (2009, 298).

2. AR theorists' case against MBR

So far I have introduced the main reasons for AR theorists' concerns with the external validity of findings in MBR. After this presentation, I will now introduce the attack mounted by AR theorists in more detail. I will begin by unpacking the AR theorists' claim that the irrationality discovered in the lab is due to the use of experimental formats that are unrepresentative of the real world. On closer inspection, it seems that the attack relies on two main claims:

a) The experimental settings used by researchers in MBR are not representative of the target system.

b) The difference between target and laboratory system is causally relevant.

To be fair, it would be misleading to present AR theorists as the only or first researchers to question the robustness of findings in MBR and to commit to a) and b). For instance, within the literature on economics, studies in MBR have been criticized because they used selected subjects and did not provide appropriate monetary incentives. Specifically, it is often claimed that, unless subjects are offered an incentive, their responses will not represent what they would do if presented with the task 'for real' (Wilcox 1993; Harrison 1994).[31] It is worth mentioning, however,

[31] Incentives are thus generally used to help emulate real-world decision-making: it is claimed that people's performance in experiments often fails to represent their competence because they are unwilling to do their best when they are insufficiently compensated for doing so. For instance, in some cases it is argued that incentives alter what the agent perceives to be her goals, at other times

that many experiments have been conducted to assess the impact of monetary incentives, showing that the biases remained and that the field of economics was required to explain them (cf. Grether and Plott 1979). Specifically, an important result was that incentives never eliminate anomalies, although they are more likely to decrease than to increase them (Camerer and Hogarth 1999).

Also AR theorists have stressed the importance of monetary incentives on occasion. For instance, in a much-discussed paper, AR theorist Ralph Hertwig and his colleague Andreas Ortmann (2001) suggest that also psychologists should use incentives whenever possible, because 'the benefits of being able to run many studies do not outweigh the costs of generating results of questionable reliability' (394). Yet, the AR theorists' main concern is with the format of the information used in psychological studies, and here I will focus on their work on the use of natural frequencies in probability judgments more specifically, since this is arguably one of their most famous and celebrated empirical contributions to research on judgement and decision-making.

One important caveat here is that I will not discuss all of the arguments offered by AR theorists that question the external validity of studies on biases. For instance, experimental research has shown that people tend to express confidence in their judgments that exceeds the accuracy of those judgments, but AR theorists have

---

that incentives induce the agent to think longer or harder (for a critical discussion of these views, see Read 2005).

claimed that people's reported *overconfidence* is due to an inaccurate selection of items, which turn out to be unrepresentative of common contexts.[32] Suffice it to say that, whilst AR theorists have offered other arguments to question the external validity of findings on human irrationality, their research on the ameliorative effect of switching from single probability formats to natural frequencies, which I discuss in this chapter, represents their flagship achievement. However, as I will emphasize below, it can also be argued that some of the conclusions I draw in the present chapter apply beyond the case of "frequency effects" and have more far-reaching implications.

With this clarification in mind, let us now present in detail the empirical case mounted by AR theorists. Specifically, these scholars have argued that recognizing the distinction between single-event probabilities and frequencies 'unearths the reasonableness hidden by the perspective of the *heuristics-and-biases* program' and allows for making 'several apparent cognitive illusions disappear' (1994, 141-2). Simply stated, 'How many subjects who test positive for the disease do actually have the disease?' is an instance of communication in terms of frequency formats. In contrast, 'What are the chances that a subject found to have a positive result actually has breast cancer?' is an example of communication in terms of probability formats. Now, with regard to claim a), AR theorists stress that MBR scholars have studied people's probabilistic reasoning using single-probability formats instead of natural

[32] For instance, AR theorist Gigerenzer has criticized the questions used to explore people's overconfidence. According to him, 'if the general knowledge questions were a representative sample from the knowledge domain, zero overconfidence would be expected. However, general knowledge questions typically are not representative samples from the domain of knowledge, but are selected to be difficult or even misleading' (1993, 304).

frequencies formats. The latter are representative of the "real world", as we are usually provided with information about risk in a frequency format. On the contrary, single probability formats are not representative of the human environment. With regard to b), AR theorists stress that violations of the norms of SPR are mainly due to the use of probability formats: when these are replaced, people turn out to follow the norms of SPR.[33]

I will now focus on two biases, often presented in MBR as indications of irrationality, whose external validity has been criticized by AR theorists: the *conjunction fallacy* and the *base rate fallacy*.

2.1 The conjunction fallacy

One of the most well known biases explored in MBR is the *conjunction fallacy*, which I presented in the introductory chapter. Let us briefly recall the classic scenario used to elicit the *conjunction fallacy*, called Linda's problem:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

[33] Interestingly, AR theorists claim that single-event probabilities and natural frequencies are mathematically equivalent. For instance, they stress that 'mathematically equivalent representations of information can entail computationally different Bayesian algorithms' (Gigerenzer and Hoffrage 1995, 679). While I will not discuss this claim here, it is worth noting that the view that these statements are mathematically equivalent is controversial (e.g., Hàjek 1996; 2009).

Now, rank the following claims according to their probability.

a) Linda is a bank teller

b) Linda is a bank teller and she is active in feminist movement

As we have already seen, a large majority of subjects rate the probability of the conjunction as higher than that of the single event, despite the fact that this violates the *conjunction rule* in probability theory. What AR theorists are eager to emphasize is that performance seems to significantly improve when subjects deal with natural frequencies instead of single probabilities. In particular, AR theorists appeal to the work of Fiedler (1988), who has reported that the rate of fallacies can be reduced by presenting Linda's problem with frequency formats, which are taken to be more representative of natural contexts. After presenting the description of Linda, Fiedler asked his subjects (Fiedler 1988 126):

There are 100 people that fit this description.

Rank how many of them are:

a) Bank tellers

b) Bank tellers and active in the feminist movement.

In this case, only a minority of the tested subjects answered that there would be more feminist bank tellers than bank tellers. Therefore, the use of frequency formats appears to reduce the rate of biases. The robustness of this effect, despite being variously interpreted, is usually taken for granted. For instance, Fiedler (1988)

reported 73% conjunction violations in the probability representation and 23% in the frequency representation. Moreover, Tversky and Kahneman (1983) found 58% and 25% conjunction violations in the probability and frequency tasks respectively.

2.2 Base rate fallacy

Another important bias discussed and allegedly mitigated by AR theorists is known as *base rate fallacy*. Consider the following problem from Casscells, Schoenberger and Grayboys (1978, 999) and presented by Tversky and Kahneman (1982b, 154) to demonstrate the generality of the phenomenon:

> If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?

Sixty students and staff at Harvard Medical School answered this medical diagnosis problem. Notably, almost half of them judged the probability that the person actually had the disease to be 0.95 (modal answer), the average answer was 0.56, and only 18% of participants responded 0.02, where the latter is the correct answer. A common interpretation of such findings is that people are prone to neglect information about base rates.

But what happens if we rephrase the medical diagnosis problem in a frequentist

way? In seminal studies on "frequency effects", Gigerenzer and Hoffrage (1995) and Cosmides and Tooby (1996) tried to answer this question. They compared the original problem (above) with a frequentist version, in which the same information was given:

> 1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.
>
> Given the information above, on average, how many people who test positive for the disease will actually have the disease?

Interestingly, when the question was rephrased in a frequentist way, as shown above, then the Bayesian answer of 0.02—that is, the answer 'one out of 50 (or 51)'—was given by 76% of the subjects. In numerous versions of the medical diagnosis problem, it seems that the improvement in subjects' reasoning is due to the transition from a single-event problem to a frequency problem. As AR theorists often put it, biases 'disappear' when questions are rephrased (cf. Gigerenzer 1991).

The "frequency effect" is now widely assumed to be fairly robust. For instance, Newell, Lagnado and Shanks recently wrote that 'the so-called frequency effect, that presenting probability problems in a frequency format often reduces judgmental biases, is now well established' (2007, 85). Moreover, this effect has also been taken

to have important applications in crucial domains. Notably, the perspective offered by AR theorists has seemingly offered a useful tool to help lay people and experts alike to reason the Bayesian way in the medical and legal domain. For example, physicians' diagnostic inferences were shown to improve considerably when natural frequencies were used instead of probabilities (Gigerenzer 1996a; Hoffrage and Gigerenzer 1998; Hoffrage, Lindsey, Hertwig and Gigerenzer 2000) and judges' as well as other legal experts' understanding of the meaning of a DNA match could similarly be improved by using natural frequencies instead of probabilities (e.g., Koehler 1996).

3. Do biases really disappear?

So, is the criticism offered by AR theorists convincing? Before I start to assess the plausibility of their critique of the external validity of findings suggesting human irrationality, it is important to note that, even if the argument were successful, it would not entail that findings in MBR were completely uninteresting. While AR theorists may have good reasons for emphasizing the importance of the generalizability of findings, this should not lead the reader to think that generalizability is the only legitimate goal for psychologists. The reader should consider that, as a number of theorists have pointed out, psychologists do not only care about the generalization of their results. For instance, in a famous paper, Mook (1983) made the point that generalizability is not always (and is in fact rarely) the goal of psychologists. Consider the following example offered by Mook. It is found that targeted people are judged to be more intelligent when wearing spectacles and

seen for 15 seconds; however, if they are observed during five minutes of conversation, spectacles made no difference. Adopting an applied perspective, what happens during the 15 seconds might be seen as less interesting than what happens in five minutes. However, it is still worth knowing that such an effect can occur even under restricted conditions: why should a person's wearing glasses affect our judgment of his or her intelligence under any conditions whatsoever? What this suggests, on Mook's view, is that even findings that are not highly generalizable can be worthy of interest and empirical investigation. It is nevertheless the case, however, that if the argument by AR theorists were correct it would have important implications for our discussion, since scholars in MBR have submitted to the view that biases are systematic and widespread in the real world.

With this clarification established, I would now like to draw the reader's attention to a first problem with the AR theorists' argument, namely that, although the "frequency effect" is generally taken to be fairly robust, the evidence collected from more recent studies on the influence of changing task formats is rather mixed. In fact, it now seems that the modification of the information format cannot be seen as a panacea. This was noted by Kahneman and Tversky, who point out that the AR theorists' claim that cognitive illusions disappear 'rests on a surprisingly selective reading of the evidence' and that 'systematic biases in judgments of frequency have been observed in numerous other studies' (Kahneman and Tversky 1996, 584). In particular, while the "frequency effect" seems to be quite effective in reducing the *base-rate fallacy*, the *conjunction fallacy* seems to be a more robust bias. For example, an average of 85% conjunction violations in probability problems and 81%

in frequency problems were discovered by Jones, Jones and Frisch (1995), who suggest that 'perhaps the effect of rewording problems in terms of frequencies is not as robust as Gigerenzer originally suggested' (1995, 113). Moreover, Mellers, Hertwig and Kahneman (2001) found that the frequency format eliminated the *conjunction fallacy* in only a minority of presented cases, viz. those without filler items and where 'and are' and 'who are' conjunction phrases were included (e.g. 'of 100 people, how many are bank tellers and are feminists?'). Further interesting findings were reported by Tentori, Bonini and Osherson (2004). The authors invited subjects to resolve different problems. One was the 'Volleyball problem':

> Professional Volleyball players have greatly changed in the course of the last decade. In particular, they have become younger yet taller. Women players in the first Italian division are on average taller than 1.80 m, ranging from 1.75 for some setters to more than 1.90 m for many spikers. Suppose we choose at random a female volleyball player from the Italian first division. Which do you think is the more probable?
>
> a) The woman is less than 21 years old (A)
>
> b) The woman is less than 21 years old and is taller than 1.77 m. (A & B)
>
> c) The woman is less than 21 years old and is not taller than 1.77 m. (A & ¬B)

Interestingly, Tentori, Bonini and Osherson (2004) presented also a version with frequencies:

> Suppose we choose at random 100 female volleyball players from the Italian first division. Which group do you think is the most numerous?

a) Women who are less than 21 years old (A)

b) Women who are less than 21 years old and are taller than 1.77 m. (A & B)

c) Women who are less than 21 years old and are not taller than 1.77 m. (A & ¬B)

What is particularly interesting in this study is that in this problem (as well as in other problems presented to the subjects), Tentori, Bonini and Osherson (2004) did not find a statistically significant difference in the magnitude of the bias between the frequency and probability format of the task. More precisely, in the probability format, 73% of the tested subjects failed to choose the option (A), whereas with the frequency format 63% of the subjects failed.[34] Interestingly, these results have been replicated and confirmed by Wedell and Moro (2008), who stress quite explicitly that 'shifting the focus from probabilities to frequencies did not significantly reduce conjunction errors', although 'it trended in that direction' (Wedell and Moro, 2008, 125). More recently, Erceg and Galic have also examined the occurrence of conjunction *fallacies* in Football betting and found that the use of the frequency-based task format did not reduce this type of bias. In their own words, 'the *conjunction fallacy* has proven to be resistant to the task format manipulation' (2014, 60).

Recently, the case has been made that the impact of frequency formats seems to interact with other variables, such as the transparency of the logical relation between

[34] It seems interesting to note, though, that while in Hertwig and Gigerenzer (1999) subjects were asked to give a frequency estimate, in Tentori, Bonini and Osherson (2004) subjects were required to choose the group with the highest frequency. However, identifying the top frequency option seems to require the same type of comparative operations that are required for ranking.

conjunction and conjunct (Sloman et al. 2003) and the response mode (Hertwig and Chase 1998; Sloman et al. 2003). It has thus been suggested that the "frequency effect" previously reported is not due to the information type but to other variables instead. In light of this, one could speculate that in some contexts it might be possible to enable subjects to solve the task in both probabilities and frequency contexts by manipulating some variables. For instance, Girotto and Gonzalez (2001) examined subjects' ability to judge posterior probability, reporting that subjects seemed to be able to make correct probability judgments in conditions in which they could easily represent all the pieces of information in the same set of possibilities.[35] The study by Sloman et al. (2003) is more relevant to our present discussion. Interestingly, researchers in this study examined the impact of different variables on violations of the *conjunction rule*, reporting better performance when filler items were introduced and when subjects were asked to rate instead of ranking.

The issues I have touched upon here are important but intricate, and they point to open empirical questions about the nature and scope of the so-called "frequency effect". I will not discuss these issues any further, but I would nevertheless like to make a quite general point: as already clarified by Kahneman and Tversky (1996), conditions exist under which the correct answer is made more transparent. However,

[35] It is interesting to note, however, that Girotto and Gonzalez (2001) introduced a representation in terms of number of chances as a way of translating natural frequencies into a language that looks like single-event probabilities. Their method has been the target of criticism, though. For example, consider the words of Hoffrage, Gigerenzer, Krauss and Martignon (2002), who point out that 'it is confusing that number of chances are called probabilities throughout the paper, because unlike probabilities, these are not single numbers in the interval [0,1] but natural frequencies disguised as probabilities' (350).

this does not mean that the biases are not robust. In particular, with regard to the *conjunction fallacy*, the contexts in which it occurs are rather numerous. Conjunction errors have been documented with various kinds of subjects, such as children (Fisk and Slattery 2005), students enrolled in different kinds of programs, and statistically sophisticated individuals (Tversky and Kahneman 1983). Furthermore, they appear in several different tasks, such as choice (Tentori et al. 2004; Wedell and Moro 2008), ranking (Kahneman and Tversky 1983; Sloman et al. 2003), betting on events (Bonini et al. 2004), and in different estimation procedures (Wedell and Moro 2008; Nilsson and Anderson 2010). What seems important for the purpose of my reply is that the findings previously discussed show that, in a number of contexts, subjects seem to be prone to *conjunction fallacies* in both contexts (namely, under the frequency format condition and the single-event condition), resulting in problems for the external validity argument suggested by AR theorists. Further research is needed if we are to reach a conclusive verdict on the issue. But since AR theorists' claims about the robustness of the effect rest on an open empirical issue, their validity cannot be taken for granted.[36]

4. Are the formats representative of the real world?

Having shown that key biases discussed in MBR do seem to be generalizable, I would now like to discuss a second important problem with the AR theorists'

[36] In addition, a remark is in order here: the discovery of even significantly different performance in the two conditions is not sufficient to support the claim that cognitive illusions 'disappear' (Gigerenzer 1991). Specifically, it is hard to concede that biases 'disappear', as long as one person out of three still tends to commit such biases.

argument. In fact, even if the "frequency effect" were as robust as AR theorists claim, it would still be possible to reject the implications of their argument. Specifically, there is another important reason why the AR theorists' argument does not seem to be as compelling as its advocates suppose. In particular, for the argument to go through, AR theorists need to show that single probabilities are not representative of the real world, while natural frequencies are. But this is problematic, or so I contend.

AR theorists do not even try to provide support for the claim that only natural frequencies formats are representative of the real world, as they seem to take this to be quite obviously the case. However, the charge that single probability formats are unrealistic seems difficult to hold. One author who has criticized this assumption most clearly is cognitive psychologist Keith Stanovich, who has researched judgement and decision-making in the *heuristics-and-biases* tradition. On Stanovich's view, much of what we currently know about the world does not come from our perception of actual events, but rather from abstract information processed and condensed into symbolic codes such as probabilities, percentages, tables and graphs. He states explicitly that:

> Banks, insurance companies, medical personnel, and many other institutions of modern society are still exchanging information using linguistic terms like probability and applying that term to singular events. (2004, 136)

Stanovich's point seems correct: we do often encounter information expressed in

terms of single probabilities. The medical domain is certainly a case in point. For example, people are typically informed about the probability that someone will develop a disease given a positive test using single probabilities. It seems hard to deny that, at least to a significant extent, single probabilities are part of our "real world".[37]

It is also worth mentioning, however, that for AR theorists the view that probability formats are unrepresentative of realistic environments does not seem to be viable. This comes out quite clearly when we notice that, according to AR theorists, single-event probabilities are not an effective way of communicating statistical information: AR theorists have blamed statistical innumeracy for human error, suggesting an ameliorative project to improve statistical reasoning by replacing problems formulated in terms of probability with problems in terms of frequencies. According to AR scholars, single-event probabilities can create significant problems when they are, for example, used by medical organizations to communicate the risks of some treatments (Gigerenzer 2002) or by an expert witness required to explain DNA evidence in court (Koehler 1996). As such, the AR theorists' proposal is to replace probability problems with frequency problems, suggesting that doing so will foster statistical reasoning and help us avoid many errors.[38] However, since AR

[37] Some scholars have complained that the notion of "real world" is rather obscure. For instance, Hammond writes that 'Gigerenzer and Todd spoil their discussion of ecological rationality by using the journalistic term "real world" in place of a meaningful theoretical term describing the ecology, or the environment' (2007, 220). Here I will preserve AR theorists' use of the term "real world", acknowledging, however, that this is by no means fine-grained.

[38] Curiously, Lee claims that, while 'the *heuristics-and-biases* project identifies the tasks in which human reasoning needs to be improved', AR scholars 'identify the conditions that actually improve and debias judgments' (2008, 64). This presentation seems misleading, since Richard Thaler and other scholars within the *heuristics-and-biases* framework have suggested a list of de-biasing methods (e.g.

theorists stress the importance of translating probability formats into frequency formats, an implicit rationale must be that in real-life situations people have to cope with probability formats: indeed, if subjects did not cope with probability formats an ameliorative project would not be needed. Therefore, probability problems *are* representative of the environment after all, and their claim that the probability format is not a feature of the environment is not feasible for AR. Here I am simply pointing to an inconsistency in the AR project, and this is an *ad hominem* argument: they cannot hold both that MBR designs are unrepresentative and that extensive de-biasing is required.[39]

5. Into the wild: in search of field data

Thus far, I have shown that the "frequency effect" is not as robust as AR theorists suppose and that there are reasons to consider single probability formats as representative of the "real world". In this final section, I will stress instead that, in order to address AR theorists' concerns, other evidence ought to be analysed, which has so far been overlooked by AR theorists. Specifically, my third line of reply runs as follows: while it is true that MBR has grown in the lab—so to speak—researchers

---

Thaler and Sunstein 2008). Yet, for some discussion of possible differences between the approaches towards de-biasing offered by AR theorists and those by scholars in the *heuristics-and-biases* tradition, see Grüne Yanoff and Hertwig (forthcoming).

[39] Before considering a third and last line of criticism of the AR theorists' argument, I would like to stress that there are other interesting directions of research on the mitigation of biases that I cannot discuss here. For instance, for a study on the effect of reputation concerns on biases see Devetag et al. (2013). In particular, however, recent research on group decision-making seems to suggest that groups are more likely to make decisions that follow the norms of SPR, while individuals alone are more likely to be influenced by biases (Charness and Sutter 2012). However, I believe that the reply I have articulated in this section might be applied to such arguments as well: if someone were to stress that people's reasoning improves when they reason in groups, we could reply that many important decisions are still individual, such that the research has only limited relevance.

have also found evidence of people's irrationality outside the lab. In particular, field experiments are investigations carried out in natural environments, rather than in laboratories: evidence coming from these experiments might be useful in assessing whether biases do actually occur in the real world.[40]

Here it should be noticed that the label "field experiment" is better understood as an umbrella term that includes different situations. For example, Harrison and List (2004) suggest a taxonomy of experiments that is useful for appreciating the variety of studies that come under the label "field experiment". While "artefactual experiments" differ from laboratory experiments only in that they deploy non-standard subjects, "framed field experiments" do incorporate important elements of the naturally occurring environment as well. However, the most interesting situations are connected to "natural experiments", in which subjects naturally undertake certain tasks and do not know that they are participating in an experiment.

It is important to stress, here, that even seminal studies in MBR have appealed to field experiments. To appreciate this, consider that Tversky and Kahneman (1983)

[40] It is true, on the one hand, that field experiments are not necessarily representative of frequent situations. Consider, for instance, Blavatskyy and Pogrebna's (2010) field experiment, which was based on the television show *Deal or No Deal*, where a contestant is endowed with a sealed box containing a monetary prize between one cent and half a million euros. In the course of the show, the contestant is offered the chance to exchange her box for another sealed box with the same distribution of possible monetary prizes inside. This scenario offers a unique natural experiment for studying *endowment effects* under high monetary incentives, where such effects refer to people's tendency to set a significantly higher value for an object if they actually own it than they would if they did not own it (Thaler, 1980). But contexts like these might bear only little resemblance to those we experience on a daily basis. This is a point that we should not ignore. At the same time, it is also true that, generally speaking, field experiments seem to offer powerful and privileged tools for the investigation of people's behaviour in the real world.

themselves combined laboratory data and field data in their seminal work, examining for example the intuitions of expert physicians. Remarkably, the evidence provided suggests an occurrence of *conjunction fallacies* in natural settings as well. Tversky and Kahneman asked practicing physicians to make predictions on the basis of some clinical evidence. Subjects were given problems of the following type:

> A 55-year-old woman had pulmonary embolism documented angiographically, 10 days after cholecystectomy. Please rank in order the following in terms of the probability that they will be among the conditions experienced by the patient (use 1 for the most likely and 6 for the least likely). Naturally, the patient could experience more than one of these conditions.
>
> dyspnea and hemiparesis (A&B)
>
> syncope and tachycardia
>
> calf pain
>
> hemiparesis (B)
>
> pleuritic chestpain
>
> hemoptysis

What is interesting to note is that physicians appeared prone to make reasoning errors. Specifically, dyspnea was considered by the physicians to be representative of the patient's condition, whereas hemiparesis was judged as very atypical. The conjunction of an unlikely symptom with a likely one was thought to be more likely than the less probable constituent. Therefore, this finding bears on the question of whether people are prone to commit biases in the real world: it presents evidence that the *conjunction fallacy* occurs not only in the laboratory, but in the field as well.

Still, an objector might argue at this point that the greatest problem with the abovementioned example is not whether it is presented in terms of probabilities, frequencies, or rank ordering; the problem is that the entire task of ranking propositions in terms of likelihood is highly artificial: medical doctors are normally not incentivized to state the most likely event, for then in practice they would always have to say 'I do not know'. They are compensated, in terms of pride and reputation, as well as in a more material sense, for "getting it right", and they therefore try to say more than they need to in this experiment. A number of comments can be made here. In offering this reply, the objector is moving away from the AR theorists' original formulation of the argument, as the objector would no longer be focusing the attack on the contrast between single probabilities and natural frequencies. That said, we could also offer a rather general reply at this point. Specifically, it should be noted that we have recently seen a general trend in the study of judgement and decision-making towards studying cognition and behaviour "in the wild", and this evidence seems less vulnerable to criticism than Tversky and Kahneman's seminal work. Interestingly, field experiments have gained momentum particularly in experimental economics, a field of research criticized since its birth with the claim that its findings cannot be applied beyond the laboratory context. Essentially, it is often insisted that, whereas laboratory experiments may allow for a relatively large amount of control, thus providing a high degree of internal validity, they nonetheless yield a relatively low external validity. In fact, it has recently become commonplace to complement laboratory data with data gathered via field experiments (List 2008), achieving a combination of control and realism that is not usually achieved in the laboratory (Camerer 2000; Levitt and List 2009). Data from field experiments are thus taken to

be particularly interesting and informative of behaviour in the real world.

To give an example of what this research in the wild might offer, consider recent findings on *framing effects*. Most studies on *framing effects* describe logically equivalent decision situations in either a positive or a negative light. The Asian Disease Problem presented in the previous chapter and (originally presented by Tversky and Kahneman (1981)) is a well-known example. Describing a choice between medical programs in terms of lives to be saved or lives to be lost leads to dramatically different answers, although the problems seem to be logically equivalent.[41] Just to mention an example, consider that Gächter et al. (2009) tested for the existence of *framing effects* within a natural field experiment and found *framing effects* among junior economists, but not among senior ones. In particular, the authors tested scholars' behaviour and vulnerability to bias when registering for a conference. In general, it seems that many findings suggesting human irrationality that have been reported in MBR have found or are currently gaining further support from field data. Thus, it seems that interesting evidence is available to support the claim that biases do not only occur in the lab, since people violate norms of SPR in the real world as well.

[41] While here we are not questioning that *framing effects* are instances of irrational behaviour, it is interesting to note that *framing effects* in decision-making have been claimed to be normatively defensible both in the economic and psychological literature (e.g., Bourgeois-Gironde and Giraud 2009; McKenzie 2004; Mandel 2014) and in the philosophical literature (Schick 1991; see also Bermudez 2009, Chap. 3 for a critical discussion).

6. Conclusion

In summary, this chapter has discussed and assessed the argument against the external validity of findings from MBR that has been put forth by AR theorists. According to AR theorists, many effects identified by researchers in MBR should not be considered genuine reasoning biases but rather the result of using experimental settings that are unrepresentative of the real world. This chapter took these reservations seriously and showed that they can be countered. Specifically, I have presented three main reasons for believing that this objection is somewhat exaggerated. First, contrary to what is claimed by AR theorists, in several cases frequency formats do not make biases disappear. Second, single probability formats are representative of the "real world". Third, evidence of violations of the norms of SPR comes from field experiments as well, and thus seems to be less vulnerable to AR theorists' concerns than AR theorists believe. Of course, this does not automatically license pessimistic conclusions about human rationality. It merely means that the AR theorists' external validity argument cannot be exploited in order to justify more optimistic interpretations, and that support must be found elsewhere. In demonstrating this, this chapter provides only indirect support for the views defended by MBR.

Chapter 2: Evolution and irrationality

The conclusion of the previous chapter was that people do seem to be prone to committing biases in the real world. One might claim, however, that this conclusion is at odds with evolutionary considerations: evolutionary pressures would have rendered these behaviours extinct if they happened in the real world. In particular, a number of scholars have appealed to an evolutionary argument for people's rationality (henceforth EAR), according to which we have good (evolutionary) reasons to believe that people's reasoning is not inaccurate, after all. The goal of this chapter is to show that the conclusion of Chapter 1 is not necessarily at odds with evolutionary considerations, and that such conclusion can, at least provisionally, be accepted.

1. The evolutionary argument against MBR

As we have seen in the introductory chapter, as well as in the chapter above, there is an imposing body of evidence suggesting the existence of widespread and systematic errors in reasoning and decision-making. This evidence seemingly shows that human beings are far from optimally rational. But a number of scholars have tried to oppose this view by appealing to evolutionary considerations. This should come as no surprise: it is now frequently claimed that by appealing to evolutionary theory we can push forward our understanding of key issues in the philosophy of social and cognitive sciences (e.g., Prinz and Barsalou 2000; Shapiro 2010; Schulz 2011b).

Unsurprisingly, evolutionary considerations have been offered in the attempt to better understand rational behaviour and cognition as well.

In particular, here I will focus on what I dub the "evolutionary argument for rationality" (henceforth, EAR). According to this popular argument, if organisms reasoned and made decisions inaccurately, then they would have failed to navigate the world successfully and thus would not have evolved. Therefore, reasoning errors cannot be as widespread as psychologists in MBR have supposed. EAR has been associated with several philosophers, the most prominent of which are Quine and Dennett. For instance, Dennett claims that 'natural selection guarantees that most of an organism's beliefs will be true' (1987, 75). However, it is possible to find similar points also in older literature on evolutionary epistemology. For example, Simpson pointed out that 'the monkey who did not have a realistic perception of the tree branch he jumped for was soon a dead monkey—and therefore did not become one of our ancestors' (1963, 84).

These considerations are still debated in philosophical discussions (e.g., Boudry and Vlerick 2014; Boudry, Vlerick and McKay 2015; Hazlett 2013; Rysiew 2008; Sage 2004), but Stich (1990) and Stein (1996) were the first to carefully reconstruct and assess this argument. For instance, Stein (1996) presented EAR as one of the main arguments in support of the claim that reasoning experiments cannot demonstrate people's irrationality. While there are different ways to formalize the argument, EAR might be summarized in the following way:

(1) Natural selection is the key factor driving evolution.
(2) Natural selection favours traits that increase an organism's inclusive fitness.
(3) It is more conducive to inclusive fitness to possess accurate reasoning.
(4) Hence, natural selection favours accurate over inaccurate reasoning.
(5) Hence, evolution grants that we possess accurate reasoning.

The problem, then, is that in virtue of (5) we now seem to face a conflict between the evidence collected in MBR and evolutionary theorizing. On the one hand, psychologists in MBR tell us that our reasoning is seriously flawed. On the other hand, EAR suggests that our reasoning cannot be inaccurate. There thus seems to be a puzzling contrast between evolutionary and psychological considerations. The conflict between these perspectives is expressed well by Martie Haselton and her co-workers, who write that:

> The general tendency in psychology is to interpret the supposedly incorrect judgment or reasoning as a genuine error or flaw in the mind. […] From an evolutionary perspective, however, it would be surprising if the mind were really so woefully muddled. (2009, 734)

The existence of this conflict seems be especially problematic for researchers in MBR. This comes out quite clearly once you consider that evolutionary theory is ubiquitous in the academic as well as in the popular press, and that several areas of inquiry have witnessed an especially vigorous degree of "evolutionary encroachment". In fact, allegiance to Darwinism has become a sort of litmus test for deciding who does and who does not hold a properly scientific worldview. In light of these considerations, dropping a commitment to evolutionary theorizing would seem to come at an unacceptable cost.

2. Assessing EAR

Prima facie, EAR looks quite plausible. However, commentators in the "rationality debate" have often tried to question its premises. For instance, after having reconstructed the argument, Stich explicitly takes the side of MBR psychologists, claiming that 'we are safe to assume that the existence of substantial irrationality is not threatened by anything that evolutionary biology has discovered' (1990, 70).

In particular, commentators such as Stich (1990) and Stein (1996) have contended that the premises of the argument are not in line with what we know in (and from) evolutionary biology. The idea, more precisely, is that EAR relies on an out-dated or inaccurate view of evolutionary theory. For instance, consider premise 1: there seem to be processes other than natural selection that are well known and discussed in the literature. Notably, Gregory Gibson in a book review of Wagner's *Robustness and evolvability in living systems* (2005) wrote that the book contributes 'significantly to the emerging view that natural selection is just one, and maybe not even the most fundamental, source of biological order' (2005, 237). Importantly, it is now widely acknowledged that factors like mutations and drift might have played a significant role in the evolution of traits and thus need be taken into account.

Moreover, consider premise 2: can natural selection favour traits that increase an organism's inclusive fitness? There are several problems with the view that evolution results in systems that are optimally designed. This comes out quite clearly when we

consider that optimal systems may have never been available. Because of this, appealing to the adaptive value of a system does not guarantee that this system has evolved. For instance, in the words of Stein, 'if truth-tropic mechanisms are not, in general, available, then natural selection will not be able to produce a significant percentage of them' (1996, 198).

With regard to these considerations about the availability of traits, it is important to note that these concerns are not at all untethered, as we can see by considering the literature on evolutionary biology and its focus on developmental constraints and pleiotropic effects. The concept of developmental constraint refers to a 'bias on the production of variant phenotypes caused by the structure, character, composition, or dynamics of the developmental system' (Maynard-Smith et al. 1985, 266). Interestingly, because of these developmental constraints, some variant phenotypes cannot be generated and thus are not available, even if they would have been strongly favoured by natural selection had they arisen. Minelli (2009) provides an interesting example to illustrate the abovementioned *availability problem*. Many centipede species of the genus *Scolopendra* have 21 segments, and many others have 23, but never has even a single individual with 22 segments been observed in any species. He views this general phenomenon of phenotypes with "borders" as pointing to developmental rules or laws. Understanding the rules—that is, the developmental mechanics—will enable an understanding of the evolutionary basis for the observed differences and discontinuities in animal forms. Minelli proposes an analogy to the rules of chess: a knight can reach only certain squares by moving from its current position—and these moves are the variation upon which natural selection can act.

This quite nicely demonstrates the nature of the abovementioned *availability problem*.

We should also consider the phenomenon of pleiotropy, in which one gene affects two or more distinct traits or systems. It might happen that a gene has positive effects on one system, but negative effects on another. Using a well-known example discussed also by Stich (1990, 65), let us consider the genes of albinism in arctic animals. The white coats for which these genes are responsible are adaptive. At the same time, since the genes are taken to be responsible for serious eye problems, albino animals do not see as well as their coloured conspecifics. The optimal genes would then provide albinism without bad eyes. But apparently this has never been an option that natural selection could select.

This should be enough to show that premises 1 and 2 are currently considered problematic by many evolutionary biologists. But this this does not mean that presenting considerations about the adaptive value of a trait cannot have any evidential significance at all. In fact, it can still be argued that establishing the adaptive importance of a trait provides some evidence of its evolution and existence (even though that evidence is clearly not conclusive). Moreover, it is worth noting that the abovementioned criticisms of premises 1 and 2 do not necessarily entail a departure from Darwin's theory of evolution. Specifically, while some authors do believe that the inclusion of factors other than natural selection (such as drift and developmental constraints) in the study of evolution is incompatible with Darwinian

approaches (e.g., Fodor and Piattelli-Palmarini 2010)[42], many others are more cautious (e.g., Minelli 2010), and others still refer to an "extended synthesis" (Pigliucci and Müller 2010) that encompasses several different factors. For instance, considering the relationship between developmental constraints and evolutionary theory, Sterelny claims that 'no very revolutionary shift is needed to incorporate developmental insights into an evolutionary perspective' (2000, 371). While these discussions are clearly important, what is most important for the purpose of this analysis is just to state that the first two premises are considered problematic in light of what we know from current evolutionary biology. It goes beyond the scope of this work to analyse the compatibility of this view with Darwinianism.

But there is another important point to discuss here. Notably, commentators in the "rationality debate" have generally considered premise 3 to be particularly problematic, which ultimately prevents EAR from being successful. Specifically, scholars emphasize that it is controversial whether the sort of reasoning and decision-making that maximizes the survival of the reasoner is that which best approximates reality (cf. Stich 1990; Stein 1996; Sage 2004). In attempts to debunk EAR, commentators have generally tried to make the point that inaccuracy might pay in evolutionary terms. In particular, Stich (1990) described potential cases of cognitive processes that, while adaptive, might generate false beliefs. The "Garcia effect" is a case in point. In a famous experiment (Garcia, McGowan and Green 1972), sickness was induced in rats by exposing them to radiation, after they had eaten distinctively flavoured food pellets. These rats then manifested a strong disposition not to eat food

[42] For some criticisms of Fodor and Piattelli-Palmarini's (2010) arguments, see, e.g., Pigliucci 2010; Futuyma 2010; Sober 2010, Diez and Lorenzano 2013 and Fulda 2015.

of that flavour, a disposition not to be found in rats that also ate those food pellets but were not exposed to radiation. However, these rats did not manifest a disposition not to eat food pellets of the same size and shape as those that they had eaten before they were made ill. Thus it seems that these rats associated food-related illness with the flavour of food, and not with its size or shape. These results suggest the following interpretation: these rats' cognitive faculties are at least less than perfectly reliable because, in this instance, they produced the false belief that food with such-and-such a flavour is poisonous. The rats are inferring from the fact that they got sick after eating food of such-and-such a flavour that they will get sick again if they eat food of the same flavour. The fact that rat cognition is inaccurate in this way does not entail that rat cognition does not offer an evolutionary advantage, though. Stich points out that:

> Strategies of inference that do a good job at generating truths and avoiding falsehoods may be expensive in terms of time, effort, and cognitive hardware. [N]atural selection might well select a less reliable inferential system over a more reliable one because the less reliable one has a higher level of … fitness. (1990, 61).

Importantly, it seems that two factors are notable in the case of Garcia's rats: the cost of false negatives, i.e., the risk of illness and death that comes from false negatives (Stich 1990, 62), and the cost of less fallible cognition, i.e., the 'time, effort, and cognitive hardware' (61) that would be required to maintain more reliable cognitive faculties. Thus we can speculate that rat cognition evolved in response to selection pressures and that these faculties involved in the rats' inaccurate conclusions are not only inaccurate, but also that being inaccurate, in the way they are inaccurate, makes these inaccurate faculties adaptive.

Unsurprisingly, in light of the discussion above, commentators in the "rationality debate" have typically considered EAR to be unconvincing (cf. Stich 1990; Stein 1996), because the premises of the argument neglect crucial facts about the evolution of human reasoning. In particular, the critical discussions of EAR offered by Stich and Stein are generally taken to conclusively show that it suffers from serious flaws and that the apparent conflict between EAR and MBR should be resolved by denying the conclusions of EAR. It is also worth noting that a number of evolutionary behavioural scientists have recently begun to provide further support to the conclusions drawn by Stein and Stich, showing that—and how—inaccuracy might be adaptive. A look at the literature suggests the existence of clusters of adaptive misbeliefs, systematic adaptive misbeliefs, and evolutionary tradeoffs.

In particular, evolutionary psychologists have attempted to identify cognitive faculties that might be adaptive and yet systematically generate inaccurate beliefs. For example, in the past decade Martie Haselton and colleagues have developed a mathematical theory dubbed "error management theory" (Haselton and Buss 2000). These authors argued that, in certain domains, such as representing certain aspects of the visual world, reasonable accuracy is adaptive, whereas in others it might be adaptive to systematically misrepresent the world. Biases may be directed by trade-offs in error costs. In particular, Haselton and Buss (2000) attempted to document the existence of some error-management effects. For instance, it has been reported that during brief cross-sex interaction men tend to rate women's sexual interest more highly than the women themselves did. Haselton and Buss suggested that this could

be due to an evolved sexual over-perception bias in men. They hypothesized that the fitness costs of underestimating a woman's sexual interest and hence missing a sexual opportunity were higher on average than the costs of overestimating her interest and spending time and effort in useless courtship. The same asymmetry does not hold for women's estimation of men's interest, because of women's selectiveness in mate choice (Trivers 1972) and men's willingness to engage in sex. This is just an example of the effects that evolutionary behavioural scientists have described. Nevertheless, what this suggests is that it is not at all implausible to accept that inaccurate reasoning and decision-making might be evolutionarily adaptive.

To be sure, an objector might argue at this point that we should not be too quick in drawing such conclusions. In fact, some scholars have stressed that it is crucial to explore in detail the range of circumstances in which natural selection might favour rational judgement and decision-making. For instance, Stephens argues that:

> Neither the proponents nor the sceptics have spent much time examining detailed models exploring the conditions under which evolution by natural selection would favour various kinds of beliefs and desire formation policies. Philosophers have typically answered the question about whether natural selection favours rational beliefs in a yes-or-no fashion. My contention is that this debate has been pursued at too abstract a level of analysis. (2000, 162)

To be sure, there is some truth in Stephens's remark, as it is important to quantify and qualify the claim that natural selection might result in inaccurate behaviour. As Stephens points out, more work is needed to establish and clarify the range of circumstances under which evolution might favour false beliefs. At the same time, it does not seem that this invalidates the point made before: commentators such as Stich (1990) and Stein (1996) seem to be right in claiming that EAR is controversial,

and the evidence and arguments provided so far suggest that it is at least possible that evolution might lead to the possession of reasoning systems that are inaccurate.

3. Assessing the relevance of EAR

So, one might take the outcome of the analysis presented above to support Stich's contention that 'we are safe to assume that the existence of substantial irrationality is not threatened by anything that evolutionary biology has discovered' (1990, 70). However, I would recommend being careful in drawing this conclusion. The reason why I suspect that Stich's claim is, after all, problematic, is that there is a trend in the literature to run together quite different kinds of phenomena and *explananda*. This, I think, is an even more serious problem than the one raised by Stephens and discussed at the end of the previous section.

Let me explain in detail the nature of the problem I am referring to. On the one hand, when Stich refers to psychological findings presented in MBR, he refers to behaviour that seems to violate the norms of SPR. In doing so, he is not alone: as we have seen, it is quite common to take biases as instances of behaviour that violates the norms of SPR. According to standard interpretations of MBR, what the relevant psychological findings suggest is that people are prone to violating those norms.

To remind ourselves of what violations of SPR might look like, it is worth considering, here again, a well-known instance of irrational behaviour coming from MBR, namely the violation of the axiom of *descriptive invariance*. McNeil et al.

(1982) confronted subjects with two problems concerning choices in therapy against cancer. Each time subjects were presented with a choice between surgery and radiation. But whereas the first problem was framed in terms of survival rates, the second problem was framed in terms of mortality rates. In Problem 1, the following information was given: (*A*) of 100 people having surgery, 90 live through the post-operative period, 68 are alive after one year, and 34 are alive after five years; (*B*) of 100 people having radiation therapy, all live through the treatment, 77 are alive after one year, and 22 are alive after five years. In Problem 2, the data were the following: (*A*) of 100 people having surgery, 10 die during surgery or during the post-operative period, 32 die after one year, and 66 die after five years; (*B*) of 100 people having radiation therapy, none die during treatment, 23 die after one year, and 78 die after five years. These two problems seem to contain the same information, but they frame it differently. In violation of the *invariance principle*, this difference in formulation has important consequences: whereas only 18% of the respondents choose the radiation therapy in Problem 1, 44% choose it in Problem 2. This is a clear example of the *framing effects* discussed in previous chapters. In MBR, violations of the axiom of *descriptive invariance* have been taken to represent instances of biased and irrational behaviour. Notably, what Stich (1990), Stein (1996), and other commentators have in mind when presenting key findings from MBR seem to be effects like *framing effects*, *conjunction fallacies* and the like.

On the other hand, it seems that a different criterion of accuracy is involved in discussions about EAR. More precisely, EAR seems to define accurate reasoning by appealing to the criterion of empirical accuracy: what defines and constitutes "good"

reasoning is whether people's beliefs represent the world accurately, and irrationality is thus identified with the possession of false beliefs. If I am right about this, and EAR is not crucially defined by appeal to the norms of SPR, it then seems to follow that the conclusions of EAR might not be directly relevant when it comes to assessing the compatibility of evolutionary considerations with the conclusions drawn in MBR. If researchers want to attack the conclusions of MBR using evolutionary considerations, then they should make sure that they are working with the same concepts of accuracy and rationality.

4. A new evolutionary argument?

Let us take stock of what we have shown so far. We have accomplished two main tasks. First, it has been shown that, contrary to the claims of advocates of EAR, evolutionary considerations do not guarantee that people's reasoning is accurate. Second, it has been argued that the psychological findings from MBR that are commonly mentioned by commentators on EAR do not seem to involve the same criteria of accuracy used in EAR. As a result, supporters of EAR should be careful when applying the conclusions of EAR to an assessment of the conclusions of MBR, even if EAR turned out to be convincing.

One might wonder, at this point, whether there might not be a new evolutionary argument against the conclusions of MBR that, instead of relying on the criterion of empirical accuracy, appeals to the norms of SPR. In fact, it might even be argued that some scholars sometimes associated with EAR had such picture in mind when they

expressed their evolutionary ideas about the possibility of irrational behaviour. For instance, Fodor seemed to have such norms in mind when he pointed out that 'Darwinian selection guarantees that organisms either know the elements of logic or become posthumous' (1981, 121). It is also worth noting that other theorists in the literature have expressed similar considerations. For example, it has been argued that evolution results in rational choice as summarized by the laws of logical thought (Cooper 2003).

At the same time, a number of authors have recently tried to dismantle this approach and to offer instead evolutionary considerations not to *oppose*, but rather to *account for* the existence of violations of the norms of SPR.[43] It is interesting to note that AR theorists themselves are among the scholars who have pursued this avenue. For instance, Stevens claims that:

> The frequent labelling of behaviours as irrational, anomalies, or biases assumes a particular perspective on rational norms. … A broader, evolutionary perspective cautions against using these labels, emphasizing instead an understanding of the decision goals, selection pressures, and decision-making environment. … Natural selection does not favour coherence to rational norms, but increases fitness relative to others in the population. (2008, 295)

As the reader can appreciate from this quote, according to AR theorists, evolutionary thinking does not necessarily conflict with the discovery of deviations from SPR in MBR. How is this possible?

[43] It is worth mentioning that Cooper (1989) himself offers a discussion of why a probability-matching strategy might be fitness-optimizing. More precisely, he argued that 'optimal fitness not only fails to imply behavior that is uniformly rational by classical standards, it logically entails the occurrence of behavior patterns that seem clearly irrational. Thus it is evolutionarily predictable that even perfectly adapted individuals may sometimes exhibit what could appear (classically) to be blatantly unreasonable behavior' (1989, 479). In other words, Cooper tries to show the selective advantage of a specific pattern of behavior that violates classical decision theory and, therefore, SPR.

In some cases, inaccuracy (defined here as a departure from the norms of SPR) might be directly advantageous. For instance, according to AR theorists, 'in social situations […] it can be advantageous to exhibit inconsistent behaviour in order to maximize adaptive unpredictability and avoid capture and loss' (Todd and Gigerenzer 2000, 737). For instance, natural selection may favour the evolution of counter-strategies that make an organism's actions more difficult to predict, including innate randomization mechanisms that reduce the risk of consistent behaviour and predictability.

These considerations might sound quite abstract. It is worth noting, however, that some support has recently come from research carried out, in particular, by Trivers (2011). According to Trivers, self-deception is supposed to produce incoherence, where the latter is defined as the mind holding contradictory beliefs. The goal of this form of incoherence is to outwit competitors; it thus serves an adaptive goal. It does so because the self-deceiver fails to give the cues that come with conscious deception, cues that the opponent may be able to pick up. According to Trivers, self-deception makes it more difficult for competitors to detect the intention to deceive. Specifically, Trivers's claim is that 'self-deception evolves in the service of deception' (2011, 4). The holding of contradicting beliefs (e.g., 'I never lie' and 'white lies are ok') occurs largely unconsciously and helps one achieve the successful deception of others. Trivers's (2011) theory might thus be able to account for the evolution of inconsistency between what we (honestly) believe we will do

and what we actually do. On Trivers's view, the adaptive value of self-deception outweighs its costs.

One might claim that the considerations offered by AR theorists (and supported by Trivers) help to shed light on at least some interesting psychological findings in MBR. In particular, in social psychology systematic discrepancies between what people say they will do and what they do are often labelled irrational biases. Notably, such an attitude-behaviour gap has been of intense interest to social psychologists for decades (e.g., Wicker 1969). What is of particular interest here is that AR theorists (and Trivers too) seem to appeal to different criteria of accurate reasoning than those defended by scholars commenting on EAR: while the former seem to be interested in whether evolution can lead to behaviour that departs from the norms of SPR, commentators interested in EAR have focused on whether evolution can lead to empirically inaccurate predictions.

But other considerations in support of the evolution of behaviour that violates the norms of SPR can be offered, stressing, for instance, that following the norms of SPR might be expensive in terms of time, effort, and cognitive hardware. In particular, adhering to logical consistency might come at too high a biological cost, as checking for consistency among beliefs might be extremely demanding. Since the biological resources required to check for consistency among beliefs could be used in other ways to confer immediate benefit to the organism, it might follow that natural selection favours incoherent cognitive faculties.

In addition, strategies that violate SPR might evolve as a by-product of inferences that people follow in their decision-making. For example, AR theorist Gigerenzer (2000) describes a rule that he calls "minimalist". When faced with options that differ in more than one dimension, an agent using the minimalist rule picks a dimension at random and then selects the option that ranks highest on the chosen dimension. This rule can result in intransitive choices, but it performs well when compared with rules that are more complex and that do not violate transitivity. But violations of transitivity can occur also when cues are used in a systematic order. In particular, in the strategy that Gigerenzer dubs "Take the Best," each option has a value on each of several cues. For these cues, the values may be "present" (+), "absent" (–), or "don't know". When choices have to be made between two different options, these are first compared to cue 1. If one option has a + and the other has a –, then the one with the + is chosen. If cue 1 does not result in a decision, those following the procedure attempt to make a decision on the basis of cue 2, and so on until a decision is reached. There is evidence that using this strategy might lead to adaptive behaviour (cf. Gigerenzer 2000), and this might suggest that strategies that violate SPR but lead to accurate predictions might evolve.

Considerations offered by other evolutionary behavioural scientists seem to support the case made by AR theorists and their claim that violations of the norms of SPR might be a by-product of evolutionarily adaptive behaviour. For instance, Houston et al. (2007a) suggested that violations of the axiom of *transitivity* could be a by-product of optimal foraging and result from an optimal state-dependent strategy that maximizes the animal's probability of long-term survival. Foraging options

differ in terms of mean energetic intake and risk of predation: some are safer but provide only low energy intake; others promise a high yield but at considerable risk. Depending on its energy reserve and the necessity of choosing a higher intake option as an insurance against starvation, an animal will change its preferences between the foraging options: when the energy reserves are high the animal chooses options that are safe but have lower yield; when reserves are low, the animal should take risks to procure higher intake and to avoid starvation. In the real world of animals, the objective of being transitive may thus get into the way of trying to survive. Houston et al. (2007b) make a similar argument about intransitivity in humans, claiming that it has been mistakenly interpreted as a form of irrationality, and have identified environmental structures in which violations of transitivity are adaptive.[44]

All in all, some considerations offered by AR theorists, as well as by other evolutionary behavioural scientists, might be used to *account for* the presence of violations of norms of SPR, rather than to *oppose* it. However, some important remarks are in order here.

First, one might point out that there is something odd in the evolutionary considerations offered by AR theorists. On the one hand, as we saw in Chapter 1, AR theorists claim that biases 'disappear' in realistic contexts, and thus people do not seem to violate norms of SPR. On the other hand, their abovementioned evolutionary considerations suggest that violations of SPR might have been adaptive and thus evolved. One might clearly point out that AR theorists cannot have it both ways. If

[44] For a recent review of these findings see, e.g., Fawcett et al. (2014).

pressed, AR theorists can probably refine their claims and say that they are of intermediate strength and therefore compatible. More precisely, they can reformulate these claims as saying that the occurrence of biases is not too frequent in the real world and that the existence of these biases can be accounted for from an evolutionary perspective. Here I will not engage in the detailed exegesis of AR theorists' work that would be necessary to clarify whether such a move is available in light of their claims or stated commitments. We have already shown that biases do not disappear in the real world, and what is interesting to note here is that AR theorists' appeal to evolutionary considerations in order to account for findings in MBR opens up new directions in the study of the links between evolution and rationality, suggesting ways to account for behaviour that violates SPR from an evolutionary perspective.

Second, one might argue that such evolutionary considerations cannot really account for the occurrence of biases, as they are only sketchy and represent nothing more than speculations on the adaptive value of strategies that depart from the norms of SPR. To be sure, there are several important limitations in such evolutionary considerations: as we have argued in the previous sections, we need to be careful when drawing conclusions about the existence of traits from their alleged adaptive value. As it turns out, evolutionary behavioural scientists have often been charged with drawing unwarranted conclusions about the existence of traits from considerations about their potential adaptiveness, and with using an inappropriate methodology (but see Machery and Cohen 2012 for some critical discussion of these charges). This is an important caveat. But this should not prevent us from accepting

the rather modest conclusion that I seek to defend in this chapter. Specifically, what I want to show here is just that there are no obvious evolutionary reasons that undermine the conclusions drawn in MBR. In fact, as we have seen, evolutionary considerations have been presented in the literature not only to oppose, but also to account for the existence of behaviour that violates the norms of SPR. What I wish to conclude is that critics of MBR have not as yet offered a knockdown argument based on evolutionary considerations that can undermine the view that people are prone to biases. Until critics of MBR come up with a clearer and stronger argument based on evolutionary considerations, it is not unreasonable to accept the conclusions of MBR, defended in Chapter 1, suggesting that people are guilty of systematic biases. This, however, is but a provisory conclusion, and I suggest that more work should be undertaken on these topics.[45]

5. Does this speak against MBR, then?

One might still reply, however, that here I am missing a crucial point, and that I am simply misunderstanding the relevance of these evolutionary considerations. An objector might in fact ask the following question: if strategies that violate then norms of SPR were shown to have evolved because they were adaptive, would not this still count as evidence against MBR? After all, one might be tempted to follow Davis Sloan Wilson, who points out that:

[45] I will come back to the claims I make in this chapter. In particular, in Chapter 6 I discuss whether research on individual differences in judgement and decision-making is at odds with an adaptationist perspective. Moreover, in chapter 7 I will question the idea that biases are best characterized as violations of the norms of SPR.

> Rationality is not the gold standard against which all other forms are to be judged. Adaptation is the gold standard against which rationality must be judged, along with all other forms of thought. (2002, 228)

There seems to be some tension here. On the one hand, as we have seen in the previous sections, one could take the view that evolution might favour violations of SPR as evidence supporting the conclusions of MBR that we presented in Chapter 1: what these evolutionary considerations suggest is that it is not at all inconceivable that biases occur in the "real world", contrary to what some critics of MBR suggest. But one might also take these evolutionary considerations as evidence against the conclusions of MBR: on this view, showing that behaviour that departs from SPR is evolutionarily adaptive speaks against the normative force of those norms of rationality.

Wilson's quote is representative of a trend: there are a number of scholars who take evolutionary arguments to have far-reaching normative implications. For instance, a clear expression of this approach has been offered by Cooper, who believes that, since tenets of SPR and evolution can be shown to pull in opposite directions, 'the traditional theory of rational choice is invalid as it stands, and in need of biological repair' (1989, 479). This is a bold claim and the question that arises at this point is whether considerations about the adaptive value of behaviour that departs from SPR speak against the normative force of the latter. Careful analysis is needed here. In fact, while there are indeed scholars who take evolutionary arguments to have far-reaching normative implications, other theorists have issued warnings against deriving such normative conclusions. Take, as an illustration, Gilboa, Postlewaite and Schmeidler, who claim that 'evolutionary arguments cannot

serve as definitions of rationality per se' and that 'it is when one confronts novel situations […] that the evolutionary arguments appear the weakest' (2012, 27). These are important points.

First, whether evolutionary considerations bear on the validity of norms of SPR seems to depend on how we define rational behaviour and how we justify rational norms. Notably, while AR scholars define as rational behaviour that achieves an organism's goals, scholars such as Gilboa claim that 'a mode of behaviour is rational when a person is not embarrassed by it, even when it is analysed for him' (2010, 5). Moreover, while AR scholars justify normative claims by appealing to the consequences that different strategies lead to in the "real world", other scholars reject pragmatic justifications and claim instead that the justification of norms is guided by reflection and arises directly through human intuition. However, it is important to emphasize that, as we saw in the introductory chapter, it is quite common to take considerations about whether a particular reasoning strategy or set of strategies is conducive to success to bear on normative issues: this is a widely held commitment, and, importantly, it is generally shared by both AR theorists and researchers in MBR.

Second, even if we accept that success matters for the justification of norms of rationality, we have to be careful when drawing normative conclusions about the validity of SPR based on evolutionary considerations. In particular, scholars such as Stanovich seek to reject the relevance of evolutionary considerations by arguing that we should not overlook the important distinction between evolutionary adaptation

and instrumental rationality (utility maximization given goals and beliefs) (cf. Stanovich and West 2003; Stanovich 2004). In Stanovich and West's words:

> The key point is that for […] means/ends rationality, maximization is at the level of the individual person. Adaptive optimization […] is at the level of the genes. In Dawkins's terms, evolutionary adaptation concerns optimization processes relevant to the so-called replicators (the genes), whereas instrumental rationality concerns utility maximization for the so-called, which houses the genes. (2003, 660)[46]

This is certainly a sensible point: what we need to show to make an interesting case against the normative value of SPR in an assessment of behaviour and cognition is that reasoning strategies violating the norms of SPR can also be successful in current environments and in the achievement of people's personal goals. More precisely, even if we accept that strategies violating norms of SPR can be evolutionarily adaptive, we then need to show by other means that strategies violating such norms can be successful even in solving typical contemporary problems and for the achievement of personal goals. In other words, what we need to show is that biases are not detrimental to the achievement of people's instrumentally rational behaviour.[47]

---

[46] I would like to emphasize that, on occasion, scholars such as Stanovich seem to fail to distinguish between two different objections to AR: i) that AR researchers have focused on the environment of evolutionary adaptedness instead of current environments, and ii) that AR researchers have focused on evolutionary and not personal goals. For instance, Stanovich and West write that 'despite their frequent acknowledgements that the conditions in the EEA do not match those of modern society, evolutionary psychologists have a tendency to background potential mismatches between genetic interests and personal interests' (2003, 172). Notably, one could focus on the attainment of fitness enhancing behaviour in current environments, since, after all, natural selection is still a force steering our evolution.

[47] These considerations are particularly important, as a common concern is that rational behaviour cannot arise from the application of evolutionarily adaptive heuristics, since they are tailored to very specific environments and contexts, and are likely to misfire in new and different contexts. As an objector might point out, the hallmark of rationality is in fact the ability to deal with novel and complex situations, and a truly rational agent should be able to operate successfully in a wide variety of environments. This is an important issue. It should be noted, however, that AR theorists have argued on several occasions that our heuristics are robust and successful in a wide range of modern

This does not mean that evolutionary claims are completely irrelevant to such normative considerations. In fact, it seems that evolutionary considerations can still be used as a heuristic device to make helpful suggestions about which contexts of choice and decision-making are worth exploring further, and to point to some overlooked phenomena that it would be helpful to know more about. However, what ultimately matters when assessing the validity of the norms of SPR *qua* benchmarks of rational behaviour and cognition is whether following such norms is conducive to successful behaviour in contemporary environments and for the achievement of personal goals. For these reasons, scholars should be very careful when drawing normative conclusions from the evolutionary considerations discussed in the previous section.

6. Conclusion

In this chapter, I have explored whether evolutionary considerations can be used to oppose the conclusions drawn in MBR. I presented a popular evolutionary argument against MBR—that I dubbed EAR—and discussed why its premises seem to be controversial. I have also shown, however, that it is unclear whether the conclusions of EAR are relevant to the verdicts of MBR, given that the criteria used in discussions around EAR seem to be different from those used in discussions around MBR. I then pointed to recent work on the link between evolution and rationality and suggested that there exist other evolutionary arguments in which evolutionary

environments as well. For considerations on the robustness of AR theorists' suggested findings on the performance of heuristics, the reader should focus on the discussions offered in Chapter 3 and Chapter 7.

considerations seem to be used not to *oppose*, but rather to *account for* the existence of behaviour violating the norms of SPR. Finally, I concluded by arguing that normative conclusions do not automatically follow from the evolutionary considerations presented above.

Chapter 3: Blame it on the norm!

In Chapter 2, I showed that there is no compelling evolutionary reason to reject the idea that people are prone to systematic and widespread biases. Yet, in discussing how biases could be evolutionarily adaptive, I also raised the question of whether biases could be "rational". The goal of this chapter is to further address this question. More precisely, in this chapter I seek to defend AR theorists' contention that behavior violating the norms of SPR can be successful and adaptive. In light of this, I stress that evaluating reasoning and decision-making according to SPR is normatively problematic, and I challenge the view that SPR provides us with universal benchmarks of rationality.

1. MBR and SPR

As the reader will recall from our introductory chapter, scholars in the field of judgment and decision-making typically justify the norms of SPR by appealing to their being conducive to success in the real world. More precisely, most psychologists in MBR seem to contend that such norms have normative force only if and as long as following them is conducive to successful behavior and cognition. The reader might recall, for instance, the words of Jonathan Baron, president of the *Society for Judgment and Decision-Making*. According to him:

> If it should turn out that following the rules of logic leads to eternal happiness, then it is rational thinking to follow the rules of logic (assuming we all want happiness). If it should turn out that, on the other hand, carefully violating the laws of logic at every turn leads to eternal happiness, then it is there violations that should be called rational. (2000, 53)

Researchers in MBR are generally quite convinced that violations of SPR are conducive to maladaptive behavior. For instance, Milkman et al. point out that 'massive costs can result from suboptimal decision-making' (2009, 379). But these statements are then susceptible to empirical scrutiny. More precisely, we should ask the following question: is it really the case that following the norms of SPR leads to successful behavior, as scholars in MBR are eager to point out? As we will see in the remainder of this chapter, AR theorists seek to resist this conclusion, arguing that the evidence suggesting the existence of such a link is less convincing than researchers in MBR believe, and that there is, in fact, growing evidence contradicting these statements. Specifically, AR theorists try to defend the claim that following the norms of SPR is, in a number of contexts and occasions, unnecessary for the achievement of successful behavior, and that following norms of SPR can in fact be detrimental to the achievement of successful behavior.

2. The problem of the absence of evidence

It might be tempting to think that great achievements of humanity, such as building bridges or computers, rest on the application of standard principles of rationality, and that this shows, in turn, that there is a strong link between adhering to SPR and achieving successful behavior. But AR theorists seem to challenge whether the link is really so strong. In particular, they question whether there really is convincing evidence available to us suggesting that violating the norms of SPR leads to maladaptive behavior. Notably, while money pump and Dutch book arguments are often discussed in the literature, it may be argued that the threat of becoming a

money pump carries little weight if such scenarios are never instantiated. In fact, on 26 October, 2011, Gigerenzer, Hertwig and Arkes posted a request on the electronic mailing list of the *Society for Judgment and Decision-Making* (personal communication). Specifically, these researchers asked the approximately 1000 members of the society whether they knew of studies in which violations of the norms of SPR resulted in demonstrated costs. The members could generate very few examples.

One such study was that by Bonini, Tentori, and Osherson (2004). The issue in question was the *conjunction fallacy*: as the reader will recall from previous chapters, this bias is manifested by assigning a higher probability to the conjunction of two events than to one of the conjunction's two constituents. This behaviour appears to violate probability theory and, therefore, SPR. In one case discussed by these authors, participants could distribute 7 euros among three predictions in a manner that reflected their beliefs that the prediction would become true in the future. Here is their example (209):

> Thanks to new labour laws throughout Europe:
>
> a) Employment will increase by 5% (p).
> b) Employment will increase by 5% and economic growth will not be less than 2% (p-and-not-q).
> c) Employment will increase by 5% and economic growth will be less than 2% (p-and-q).

In this case, participants bet an average of 1.99 Euros on p and 3.15 Euros on p-and-q, thus indicating that they believed that the conjunction was more likely than one of its constituents. If q were wrong, however, this violation of a norm of SPR would

actually cost the participants money.[48] Thus, this study comes quite close to demonstrating the costly effects of violating the norms of SPR.

Another candidate was a study by Azar (2011), who modified a study described in Tversky and Kahneman (1981), in which people were willing to drive 20 minutes to save $5 on a $15 calculator but were unwilling to do so to save the same amount on a jacket that cost $125. According to Tversky and Kahneman, the difference in the relative amount of the savings (33% versus 4%) should not influence the money-time trade-off. According to Azar (2008), people who exhibit this standardly irrational behaviour would indeed pay more for a portfolio of goods compared to persons who did not.[49]

What this suggests is that finding clear demonstrations of actual cases of maladaptive behaviour caused by violations of the norms of SPR is not as easy a task as researchers in MBR seem to believe. Clearly, this does not imply that such demonstrations cannot be found. However, it seems that AR theorists deserve credit for attempting to raise researchers in MBR from a sort of "dogmatic slumber" and for challenging the uncritical endorsement of normative standards.

[48] Otherwise it would not cost money.
[49] One might argue, however, that this behavior does not necessarily violate norms of SPR. Specifically, it is not entirely clear why the difference ($5) should count as rational, and not the ratio ($5/$15 versus $5/$125). After all, theories of expected utility assume diminishing returns.

3. A stronger case against SPR

The problem, however, is not only that the evidence documenting the existence of a clear link between the SPR and adaptive behavior is limited, but also that, at least in some contexts and domains, behavior that seems to violate such norms seems to be nevertheless successful and adaptive.[50]

The possibility of scenarios in which adaptive and successful behavior is achieved in spite of violations of the norms of SPR had been discussed in the literature prior to the work of AR theorists. Here I will just present some of the earlier considerations. Take, for instance, Sen (1993), who suggested that behaviour that violates the axiom of *independence of irrelevant alternatives* might be successful and rational. He offered the following example. Imagine that, at a dinner party, a fruit basket is passed around. When this reaches you, one apple is left in the basket. You decide to behave decently and pick nothing (x) rather than the apple (y). Yet, if the fruit basket had contained another apple (z), you could reasonably have chosen y over x without violating standards of good behaviour. Choosing x over y from the choice set {x, y} and choosing y over x from the choice set {x, y, z} seems to breach the axiom of *independence of irrelevant alternatives*, even through there is nothing irrational about your behaviour given your good upbringing. In fact, had you suspended your

[50] It should also be noticed that considerations analogous to the ones I discuss in this chapter could also be made about game theory, although in this work I do not focus on game-theoretic approaches. In particular, in some cases heuristic reasoning seems to be quite effective also when it comes to strategic interaction. Consider, for instance, the so-called "backward induction paradox" (Pettit and Sugden 1989): in a sequence of prisoner's dilemma, I might do better following a strategy such as tit-for-tat than applying the standardly rational process of "backward induction". Moreover, it seems that for the assessment of a reasoning strategy we have to consider the goals of people: as literature on "social value orientation" reveals (Liebrand and McClintock, 1988), agents can exhibit rather different strategic profiles, and what might be rational for an individualistic agent might not be rational for a pro-social individual. It should also be noticed that also the normative framework of game theory has recently been called into question (e.g., Misyak and Chater 2014).

good manners, you would not have violated SPR. Sen (2002) concludes that the idea of internal consistency of choice 'is essentially confused, and there is no way of determining whether a choice function is consistent or not *without* referring to something external to choice behaviour (such as objectives, values, or norms)' (121-2). For our purposes here, what seems important is the suggestion that, if researchers in MBR had applied the norms of SPR when assessing behaviour and cognition, they would have mistaken sensible and adaptive behaviour for irrationality.

But Sen is not the only author to point out that violations of the norms of SPR can result in benefits on the part of the decision-maker.[51] Another example, perhaps more controversial, has been suggested by Nozick (1993, 21-25). He asks us to consider the following scenario. An individual is strongly tempted to cheat on his spouse. If he does so, he will come to regard this act as a mistake for the rest of his life. The individual ultimately refrains, and does so in part because he reflects on how much he has invested in his marriage financially, emotionally, and temporally. These investments are, of course, sunk costs. But because he allows such allegedly irrelevant considerations to influence his decision-making, he is ultimately better off than he would be if he had ignored them and succumbed to temptation.[52]

The cases illustrated above seem to suggest that, if researchers in MBR try to assess behaviour against the norms of SPR, they run the risk of mistaking adaptive

[51] I want to emphasize that this overview is not meant to be exhaustive. As it turns out, other philosophers have discussed similar problems when debating whether acting irrationally might pay off (e.g., Parfit 1984, 12).

[52] Research in MBR has shown that people often fall prey to the *sunk cost effect*, where honouring sunk costs is taken to represent a violation of SPR. It is, more precisely, a departure from the theory of rational choice, and occurs when people consider past investments of money, effort, or time when making their decisions.

and successful behaviour for irrationality. These considerations are important, and in line with the case mounted by AR theorists.[53] However, AR theorists have also tried to substantiate such claims further by providing actual empirical evidence on people's thinking and decision-making.

3.1 Fast, frugal and accurate heuristics

Let us introduce AR theorists' discussion of how behaviour that violates the norms of SPR can be adaptive. Specifically, let us begin by presenting AR theorists' research on fast-and-frugal heuristics. Human and non-human animals make decisions that go beyond the available information to make predictions about the state of the world. For instance, knowing some features of a fruit, can the decision-maker infer whether this is dangerous or not? The psychological research carried out by AR theorists has revealed that simple strategies that allegedly violate the norms of SPR can lead to remarkably successful behaviour.

To appreciate this, let us recall here the idea that human and non-human animals alike are endowed with an "adaptive toolbox" (AT) of fast-and-frugal heuristics. According to AR theorists' framework of AT, the algorithms that we use to make decisions do not use all the available information. The simple heuristics we use are non-compensatory, because only the best discriminating cue determines the inference or decision, and no combination of cues can override the decision.[54] Previous

[53] In fact, AR theorists themselves have often referred to Sen's point when articulating their challenge (see, e.g., Todd and Gigerenzer 2000, 771).

[54] Non-compensatory heuristics are taken to depart from traditional standards of rationality because they violate the compensatory weighted additive (WADD) rule, which is seen as a component of some

research in the field of judgement and decision-making has revealed that people tend to avoid making trade-offs among attributes (Kahneman and Frederick, 2002, Gowda and Fox, 2002 and Payne et al., 1993).

Yet, when adopted, these strategies can be as accurate as, or more accurate than, linear models that integrate more information.[55] The *recognition heuristic* is the simplest strategy within our AT. This strategy exploits the fact that people have a remarkably effective recognition memory, and make inferences about a criterion that is not directly accessible to the decision maker, based on recognition retrieved from memory. More precisely, the yes/no recognition response is used here as a frequency estimation cue: if one of the two alternatives is recognized and the other is not, it should be inferred that the recognized alternative has a higher value. In a famous experiment, Americans and Germans had to find out which was the more populous city between San Diego and San Antonio (Gigerenzer and Goldstein 1996). Many Germans recognized the former but had never heard of the latter, and all of them chose the former over the latter, potentially relying on the *recognition heuristic*. All the Germans gave the right answer and performed better than the more knowledgeable Americans (who recognized both the cities and, therefore, were not able, therefore, to use the *recognition heuristic*).

---

versions of utility theory (Keeney and Raiffa 1976). WADD considers the values of each alternative on all of the relevant attributes and considers all the relative importance or weights of the attributes to the decision-makers.

[55] It is worth noting, however, that also other researchers before AR theorists had stressed that the use of non-compensating decision rules could result in a rational process (Payne et al. 1993). In their seminal work on adaptive decision-making, Payne et al. collected evidence that people tend to select heuristics in an adaptive way.

Apart from the fact that these strategies seem to be non-compensatory, what is of particular interest here is that variants of these fast-and-frugal heuristics can also violate transitivity, which is a pillar of SPR. The reader might want to recall from our previous chapter that AR theorist Gigerenzer (2000) has described a rule that he calls "minimalist". When faced with options that differ on more than one dimension, an agent using the minimalist rule picks a dimension at random and then selects the option that ranks highest on the chosen dimension. This rule can result in intransitive choices, but it performs well, when compared with rules that are more complex and produce transitive decisions. Now, selecting a cue at random might not seem realistic, but violations of transitivity can occur when cues are used in a systematic order. In the procedure that Gigerenzer calls "Take the Best", each option has a value on each of several cues. For each cue, the possible values are 'present' (+), 'absent' (–) or 'don't know'. When a decision has to be made between two options, they are first compared based on cue 1. If one option has a + and the other has a –, then the one with the + is chosen. If cue 1 does not result in a decision, those following the procedure attempt to make a decision on the basis of cue 2, and so on until a decision is achieved. These strategies seem to perform quite well (Gigerenzer and Goldstein 1996), and it seems that the main lesson to be drawn from AR theorists' research is that non-compensatory heuristics that sometimes violate transitivity can be as good as (and even better than) standardly rational methods in many contexts.

Yet, the world that organisms inhabit is also a social world, and the success of an organism also depends on the quality of the interaction with other organisms. It is widely believed that reliance on heuristics is particularly problematic in social contexts. In particular, philosopher Kim Sterelny has expressed such reservations by pointing out that:

> It is no accident that the examples of such heuristics in action ignore interactions with other intelligent agents, especially competitive agents. For it is precisely in such situations that simple rules of thumb will go wrong. […] Catching a ball is one problem; catching a liar is another. (2003, 53)

Sterelny's argument has been well received in the literature (e.g., Stanovich and West, 2003; Buller, 2005, 158-160; but see also Hurley 2005), but Sterelny's claim seems to be off track, not only because AR theorists have considered the nature and performance of social heuristics as well (e.g., Hertwig et al. 2013), but also because, at least in some social contexts, heuristics that violate the norms of SPR seem to lead to successful interactions with others.[56]

[56] In the main, Sterelny claims this because he assumes that these heuristics rely on information that is subject to deception, since other agents, unlike nature, constantly try to outwit us. Moreover, social environments, unlike natural environments, seem to be unstable and to change quite frequently. Since social heuristics are supposed to be tailored to rather specific environments, it is believed that they are more likely to misfire. Discussing this argument in detail falls beyond the scope of this work. However, I still want to stress that there seem to be several problems with this reasoning. A quite general one is that the notion of social environment used by Sterelny is not fine-grained. In fact, there seem to be quite different kinds of social interactions, leading to quite different social environments, and different social environments present remarkably different features. Arguably, only few of these seem to be vulnerable to the considerations and reservations expressed by Sterelny. It is also the case that researchers have found that simple heuristics can be particularly useful, for instance for combining information to make a group decision (Reimer and Katsikopoulos 2004), for finding a suitable mate who is also agreeable (Todd and Miller 1999) or for assessing how common a disease is in one's social circle (Hertwig, Pachur, and Kurzenhäuser 2005).

To see this, I will first reconsider the *conjunction fallacy*. Since I already introduced this effect in the previous chapters, I will not recall the phenomenon again. What I will recall, instead, is that this phenomenon is usually interpreted as an indication of irrationality, since it violates the *conjunction rule* of probability theory, which states that the probability of a conjunction is always smaller than or equal to the probability of one of its conjuncts. However, exhibiting this behaviour is not necessarily irrational. It might be in fact a case of genuinely adaptive behaviour, and the conversational goal of being informative may have contributed to that finding (Hertwig and Gigerenzer 1999).[57] The basic point here is that subjects may be interpreting the goal of the task as a request to be as informative as possible. This seems plausible, since in normal conversation it is assumed that the speaker is cooperative and the cover story might be considered relevant for solving the problem. Hence, if participants are trying to be informative, and are thus ordering options according to their informational value given that profile, it makes perfect sense to choose a conjunction over one of the conjuncts. Usually, the conjunction is more informative than the conjuncts, and cover stories seem to be created to produce that very effect. In that case, it seems to be more informative to say that the person in the story is a feminist bank teller than just a bank teller. It follows from this reinterpretation that violating the norms of SPR might be a key condition for successful communication with others.

[57] It is important to note, here, that the point made by AR theorists which I am discussing here is different from their objection presented in Chapter 1, according to which such effect disappears in the "real world". I will not discuss here whether these objections are compatible or not, and merely focus on the plausibility of the scenario described in this section.

3.3 Accountability and the Independence of Irrelevant Alternatives

But there are other reinterpretations of findings from MBR that deserve attention here. For instance, Lenton et al. (2013) have offered another interesting case of behaviour that might be successful in social contexts in spite of the violation of SPR. Consider, first, a study carried out by Sedikides et al. (1999). The authors of that study asked their participants to make a choice between two mates, Eligible A and Eligible B, where A and B were described according to the criteria of handsomeness and articulation. Specifically, A scored higher than B in terms of handsomeness and B scored higher than A in terms of articulation. Notably, whether a third option was presented or not seemed to affect the choosers' preferences. The third option was inferior to Eligible A on handsomeness and equal to A on articulation (so that A dominates the third option), and it was better than Eligible B in terms of handsomeness but worse than B in terms of articulation. Specifically, the introduction of the third option resulted in participants' preferences shifting from indifference (50:50) towards Eligible A (the mate that dominated the third option). This result is usually considered to be an instance of irrationality, because it involves a violation of a norm of SPR—the *independence of irrelevant alternatives*, in this case the dominated third option (Chernoff 1954; Fishburn 1973). However, it is unclear whether such behaviour is necessarily irrational. As Gigerenzer and Gaissmeier (2011, 471) point out, 'the goals of social intelligence go beyond accuracy, frugality, and making fast decisions. They include transparency, group loyalty, and accountability (Lerner and Tetlock 1999)'. Once such goals and factors are taken into account, it is possible to reinterpret the abovementioned findings from MBR. In particular, Lenton et al. (2013) offered an argument to resist the claim that

such behaviour is irrational by appealing to the importance of the goal of "accountability". As the authors of this study point out, consumer research suggests that people choose the option that dominates the third option, partly because the presence of such options makes the attribute on which the third option is lacking (i.e., handsomeness) more salient to the decision-maker. Increasing salience of the dominating cue may result in people selecting the option dominating the third one because they find it easier to justify this choice to others (Simonson 1989). The goal of accountability is a crucial aspect of people's social rationality, and in light of the importance usually placed on social networks and on the flow of mate-relevant information through them, it seems to be particularly important to make decisions that one can explain to the people around him. As Lerner and Tetlock (1999) point out in their discussion of the effects of accountability, choosing the accountable option may bring important personal benefits, as it can facilitate social interactions. In other words, selecting a mate based on the feature that stands out and is easy to justify might not be an instance of irrationality, but rather an instance of adaptive behaviour, once we consider that such behaviour may help achieve the goal of facilitating social interactions.

3.4 Wrapping up

By looking at recent developments in the study of reasoning and decision-making, we have found support for the idea that, if practitioners in MBR rely on norms of SPR when assessing judgement and decision-making, they run the risk of mistaking adaptive behaviour for irrationality, and that evaluating reasoning and decision-

making according to SPR can therefore be, and in a number of contexts is, normatively problematic. On the whole, this looks, so far, like the most serious and potentially damaging objection to research in MBR.

4. Objections and replies

Before this conclusion is fully accepted, however, it is important to address a few central objections concerning the existence, interpretation and implications of the findings presented in this chapter.

4.1 Challenging the empirical premises

A first objection goes as follows: the empirical results on which AR theorists seem to base their challenge are implausible. More precisely, the objector might argue that, even if the findings offered by AR theorists could, in principle, have far-reaching implications for the study of human rationality, the evidence and considerations they provide are controversial.

For instance, consider that, in the case of Linda's problem, it might be argued that there is no compelling evidence showing that subjects are trying to be relevant (cf. Moro 2009, 17). Moreover, one may also challenge the evidence suggesting that fast-and-frugal heuristics are empirically accurate in spite of the violation of the norms of SPR. Notably, even the application of the *recognition heuristic* discussed above, which represents a flagship achievement of the AR theorists' program and is the simplest rule in our "adaptive toolbox", looks controversial. Specifically, in today's

terms, all the Germans would be wrong: according to the 2010 US census (http://2010.census.gov), San Antonio is a more populous city.[58] So, one might ask, if the AR theorists' claim does not stand in any individual case, how can it be plausible?

These concerns are important, but it does not seem to me that they scathe the challenge from AR in any particularly serious way. In the main, this is because the objection looks plausible on a case-by-case basis, yet it is less attractive from a broader perspective. For instance, AR theorists have presented an imposing body of evidence for the claim that fast-and-frugal heuristics can be adaptive in spite of violations of transitivity and cycles of preferences (for an extensive review of such findings see Gigerenzer and Gaissmeier 2011; for some formal analyses of these results see Arlò-Costa and Pedersen 2011; 2012). More importantly, however, the conclusion of both our analysis and the claim made by AR theorists seem to be supported by other studies and bodies of literature too.

Consider, for instance, a hypothesis recently put forward by Mercier and Sperber (2011). According to the authors, human reasoning abilities are better suited to argument and persuasion than to analysis and consequential thought. Specifically, these scholars suggest that human reasoning has an argumentative function, allowing

[58] In addition to the question of whether simple heuristics can be successful, there is also the question of whether people really do use non-compensatory heuristics in their reasoning and decision-making. With regard to the second question, a number of scholars have argued against the claim that non-compensatory simple heuristics are pervasively used (e.g., Newell 2005; Hilbig 2010). As I stated in the Introduction, in this thesis I will not discuss in great detail the range, nature and frequency of use of the heuristics described by AR theorists. What is important to stress, however, is that it is generally accepted in the literature that, at least in a number of contexts, people do rely on non-compensatory simple heuristics.

us to win arguments better than we pursue truth. These considerations prompted Mercier and Sperber to reinterpret psychological findings such as those concerning the so-called *confirmation bias*, which consists in 'the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand' (Nickerson 1998, 175). According to Mercier and Sperber, the *confirmation bias* might actually be a case of "social intelligence" at its best. On Mercier and Sperber's "argumentative theory", having a *confirmation bias* might actually be helpful and adaptively rational: when one is trying to convince someone, one wants to find arguments for one's side, and that is exactly what the *confirmation bias* might help one to do.[59]

But there are other studies supporting the conclusion defended above. Take, as an illustration, the work of Burns (2001, 2004) and Burns and Corpus (2004), who suggest that believing in the so-called hot hand (and hence committing the *hot hand fallacy*) might contribute to an adaptive behavioural strategy in basketball, because it leads playmakers to pass the ball to a player with a higher scoring average in a game.[60] The authors have thus reconsidered the *hot hand fallacy* in basketball, according to which a player has a better chance of success following other successful shots. Specifically, Burns (2001) used a simulation of ball allocations to two virtual players and showed that behaviour based on the hot hand belief resulted in higher

[59] According to the authors, 'when we want to convince an interlocutor with a different viewpoint, we should be looking for arguments in favour of our viewpoint rather than in favour of hers. Therefore, the next prediction is that reasoning used to produce argument should exhibit a strong *confirmation bias*' (Mercier and Sperber 2011, 61).

[60] Other scholars have also sought to offer an evolutionary perspective on these sorts of behaviour (e.g., Scheibehenne et al. 2011).

average scores for the team than when the belief was ignored.[61] In particular, allocating the ball to the hot player would result in a small but important advantage of about one point in every seven or eight games. Furthermore, Burns argued that the greater the variability in the base rate of a player's scoring performance, the greater the advantage of the hot hand belief. Burns's mathematical model assumes that playmakers cannot detect base rates directly; their belief in the hot hand provides another, indirect source of information.[62]

Here I do not want to contend that violating the norms of SPR generally or most of the time leads to adaptive and successful behaviour. This claim would not be empirically supported and would seem to be far-fetched. Instead, what I want to argue is that there is growing evidence supporting the contention that, in a number of contexts and domains, violating the norms of SPR might lead to adaptive behaviour and cognition and that, by relying on such normative perspective, researchers would mistake adaptive behaviour for irrationality. This suggests that the task of rejecting the empirical evidence in support of AR theorists' conclusions is more difficult than it might seem. But there is also another point to note here. Even if the attack on the empirical evidence offered by AR theorists proved successful, exactly what evidence provides positive grounds for the objectors' conviction that violating the norms of SPR is costly is still unclear. As we have seen, we do not have many clear empirical

[61] It is important to note, however, that while Burns (2001; 2004) claims that the *hot hand fallacy* permits successful fast and frugal judgments of the shooting percentage of individual basketball players, he does not deny that belief in the hot hand is a fallacy and does not draw normative implications from his results. AR theorists go beyond this claim, stressing rather explicitly that behavior should not be assessed against standard normative models but measured in terms of its adaptiveness.

[62] There are other interesting bodies of literature that might be relevant to the AR theorists' point, but I do not discuss them here. In particular, in formal epistemology there are a number of results showing that, for some systems of beliefs, the more coherent they are, the more likely they are to be false (e.g., Olsson 2005).

demonstrations that repeated violations of norms such as transitivity are costly, and the abstract threat of money pump scenarios (exploitation generated by intransitive choices between three or more options) carries little weight if they are never instantiated.

4.1 People are not violating SPR

At this point, the critic might offer a different objection and question the interpretation of such findings by stressing that the norms of SPR are only apparently violated in the cases mentioned above. Over the past two decades, some evidence has been offered to substantiate this objection. In particular, some reinterpretations of previous experimental tasks in MBR have been suggested. For instance, while psychologists declared people's choices in the *Selection Task* irrational because people's choices deviated from the laws of logic, other researchers applied Bayes' rule to the same problem and rehabilitated people's choices as rational (e.g., Oaksford and Chater 2007).

It is worth clarifying the nature of the task. Wason (1966) invented the *selection task*, also known as the four-card problem, to study the extent to which reasoning about conditional statements obeys *modus tollens*. He focused on the material implication P→Q, as defined by the truth table in elementary logic. In his experiments, the P and Q were substituted by some content, such as "numbers" (odd/even) and "letters" (consonants/vowels). The material implication "→" was replaced by the English terms "if … then", and a rule was introduced, such as:

> If there is an even number on one side of the card, there is a consonant on the other.
> Four cards were placed on the table, showing an even number, an odd number, a consonant, and a vowel on the surface side.

People were asked which cards needed to be turned around in order to see whether the rule had been violated. Wason assumed that the "correct" answer was given by the truth table (*modus tollens*), which means that one has to turn around the P and the not-Q card, and no others. The reason for this is that the material conditional is false if and only if P∩¬Q. However, most people selected other combinations of cards. In Wason's and other studies, selections inconsistent with the *modus tollens* have been taken to show a serious shortcoming in people's ability to reason "rationally". However, this interpretation has been challenged.

According to Oaksford and Chater (1994, 1996), the *selection task* is better understood as a problem of optimal data selection in which participants need to decide which of the four cards is likely to provide the most useful data to test a conditional rule. Based on the assumption that the required process is one of inductive hypothesis-testing rather than deductive reasoning, and on a Bayesian model of optimal data selection, Oaksford and Chater (1994) conclude that the people's common selection of the p and q cards and, by extension, their reasoning, 'may be rational rather than subject to systematic bias' (608).

Based on this and similar reconstructions, these researchers have criticized the challenge put forth by AR theorists. Chater et al. write that:

> Advocates of ecological views of rationality […] make much of the contrast between everyday human behaviour, the success of which must be judged in the context of a specific and complex environment, and abstract classical principles of rationality. […] In short, the concern is that classical principles of rationality are unecological, and hence inappropriate as standards of real-world reasoning. (2003, 69)[63]

However, in light of considerations such as those presented above in relation to the *selection task*, these scholars argue that:

> Norms of classical rationality are crucially involved in explaining why a particular behaviour is ecologically successful. Thus, we argue that classical and ecological notions of rationality are complementary, rather than standing in competition. (Chater et al. 2003, 65)

The framework in which Oaksford and Chater work is referred to as "rational analysis" and is intended to signify a radical departure from AR. More precisely, the authors tell us that:

> The project of providing a "rational analysis" for some aspect of thought or behavior has been described by the cognitive psychologist John Anderson (e.g., Anderson 1990, 1991). This methodology provides a framework for explaining the link between principles of formal rationality and the practical success of everyday rationality not just in psychology, but throughout the study of behavior. […] According to this viewpoint, formal rational principles relate to explaining everyday rationality, because they specify the optimal way in which the goals of the cognitive system can be attained in a particular environment. (Chater and Oaksford 2000, 106)

On the face of it, this might look like a convincing objection to the AR project. After all, "rational analysis" might seem to clearly exemplify recent probabilistic approaches to cognition that have gained prominence in research in cognitive science

[63] The reader should recall that, while we use the term AR, other scholars and commentators refer to the same framework using the term "ecological rationality".

(for a discussion of the virtues and pitfalls of the Bayesian approach, see Bowers and Davis 2012a; 2012b; Jones and Love 2011; for a reply see Griffiths et al. 2012). Yet, I think that this strategy suffers from a number of problems.

First, it is important to represent the claims by AR theorists correctly. If AR theorists claimed or needed to claim that following SPR is never (or rarely) conducive to adaptive behavior and cognition, then the evidence discussed by advocates of "rational analysis" would seem to offer a scathing challenge to AR. But that is not what they are claiming. What it is claimed, instead, is that the evidence in favor of the link between following SPR and adaptive behavior is less convincing than generally thought, and that, in fact, there exists some evidence speaking against such link.

Second, while the strategy adopted by "rational analysis" theorists and described above might look plausible in a number of cases, it is unclear or even doubtful whether it could be successful on a broader scale. For instance, in the Linda problem, the subject may be applying some norms of SPR to a different construal of the problem that takes into account the fact that the experimenter has presented the subject with a cover story. But in many other cases it seems that such explanations in line with tenets of SPR are more difficult to provide. For instance, in cases such as violations of *transitivity* resulting in fitness maximizing behaviour, or violations of *transitivity* entailed by accurate fast-and-frugal heuristics, this strategy might not be easily applied, and thus begins to lose some of its appeal. Here I do not wish to be dogmatic and claim that such reinterpretations cannot be offered. What I want to

emphasize, instead, is that the burden of providing them falls on advocates of "rational analysis", and that the task seems likely to be a difficult one.

Third, and perhaps more importantly, this would be at best a partial victory for advocates of the "rational analysis" movement. Recall, first, that AR theorists argue that, if scholars working on rationality are interested in assessing adaptive and successful behaviour, then applying the norms of SPR is problematic, as they would run the risk of mistaking adaptive behaviour for irrationality. But now it is clear that the considerations offered by Chater, Oaksford, and their co-workers could also be used to support the AR theorists' case against MBR and against a mechanical application of the norms of SPR. As we have seen, according to these authors, agents performing a reasoning task could be applying other norms of SPR to a different task. This would still suggest that the norms of SPR could not and should not be used as universal benchmarks of rationality: on the perspective of "rational analysis" what counts as a correct norm to apply depends crucially on the context and the goals of the reasoner. So, both AR theorists and their objectors from "rational analysis" agree that the mechanical application of the norms of SPR is problematic. They rely instead on considerations about contextual factors and on the importance of the goals of the cognizer in the assessment of rational behaviour and cognition.

4.3 Rationality does not always pay

At this point, the objector might accept both the findings presented in support of AR theorists' conclusions and their interpretation in terms of violations of the norms of

SPR. Still, the objector can nevertheless reject the suggested implications of such findings.

One way to do this is by arguing that we just have to accept that, at least in some contexts, being rational does not pay. The problem of rewarded irrationality has generally been discussed in the literature with regard to the so-called Newcomb's problem (Nozick 1969).[64] Specifically, it is sometimes acknowledged that rationality does not always pay and that there may be environments in which being rational is "maladaptive", but this consideration is taken to be unproblematic: we just have to bite the bullet. In other words, there may be environments in which being rational is "maladaptive", but if irrationality is rewarded then we can only infer that irrational people will be better off: we do not get to infer anything about rationality.

I do not find this reply particularly convincing. As we have seen, MBR researchers traditionally justify the norms of SPR by stressing that these are conducive to success. From this perspective, however, it seems hard to count as irrational violations of the norms of SPR that lead to adaptive and successful behavior. AR theorists seem to drive their point home when they argue that, given that nature of common justifications of the norms of SPR, we have no grounds for considering the apparently "faulty" yet adaptive behavior discussed above as irrational.

[64] Notably, Robert Nozick (1969) presented a dilemma for decision theory. He constructed an example involving a being that can make preternaturally accurate predictions about one's future decisions and in which the standard normative principle of dominance conflicts with the principle of expected-utility maximization.

4.4 A pragmatic defence of standard rationality

Another objection follows. We have seen an attempt to resist the implications of findings presented by AR theorists. This sought to deny the relevance of considerations about whether following the norms of SPR leads to success. The last objection I consider here continues to question the implications drawn by AR theorists, but does it in a different way. In other words, the objector now wants to resist the conclusions of AR theorists *while* accepting that the standards offered by SPR are flawed. In fact, accepting traditional normative standards in spite of the acknowledgement of their flaws is not completely uncommon. As Nozick once said that 'of course, our current standards of rationality are not perfect—in what year should we suppose they became so?' (1993: xiii).

Specifically, while the objector here accepts that considerations about which strategies are conducive to success might play a role in normative discussions, she also sees that other factors do matter for our choice of the right evaluative standards. More precisely, while in a number of contexts and domains applying the norms of SPR might result in mistaking adaptive behavior for irrationality, it is unclear whether there is a genuine alternative normative framework that we can apply to investigate rational behavior and cognition. The objector might thus want to offer here a pragmatic defense of SPR. As Wendt points out while commenting on these considerations I offer, 'giving up on a universal standard of rationality seems, in practice, to mean giving up on any normative model at all' (2015, 165). But this is problematic, as we want to leave room for the assessment of behavior and cognition.

More precisely, according to our objector, the framework of AR does not offer workable standards of rationality. Normative standards, to be workable, should allow us to assess the quality of performance. After all, people are interested in making good decisions and choices, and normative standards should allow us to assess our choices and decision-making strategies, offering tools for the assessment of decision-making as and when it happens. It seems safe to state that the norms of SPR represent workable standards, as it is shown by the fact that they can be applied to virtually any context and task, and still play a prominent role in the assessment and regulation of decision-making. From the perspective of AR, on the other hand, in order to assess rational behavior and cognition, we need to establish what counts as adaptive behavior and then assess whether the behavior exhibited by the agent is adaptive.

Sometimes, this can be done quite easily. For instance, when AR theorists asked their subjects which city was the more populous between San Antonio and San Diego, it was quite uncontroversial that a relevant criterion would be the empirical accuracy of the prediction. Moreover, by considering facts about demographics that are accessible to them, researchers could easily assess the answers given by the subjects.

At the same time, cases like this might seem ideal: the situation seems to change rather radically when we move towards more complex situations. Notice, first, that in some cases it might be difficult to identify the relevant goal. Stevens (2010, 114) argues that the analysis of decision-making carried out by AR theorists supports the

continuity between human and non-human minds, but we should nevertheless take care not to overlook differences. In the case of animal cognition it might seem plausible to describe the organism's goal as that of maximizing fitness, and assume this to be the relevant goal. However, when we move away from this context, and start to investigate human rationality through the adaptiveness of behaviour, it becomes less clear what would count as adaptive and successful behaviour. When we leave the context of animal cognition and explore human rationality, not only does the sophistication of the analysis increase, so does the difficulty of establishing what the relevant goals actually are. There are other cases, in which it is similarly difficult to assess performance against the relevant goals. AR theorists do not seem to be aware of this problem, which is probably due to the fact that their research is mainly based on contexts where the assessment of behaviour is rather straightforward. In particular, a great deal of their empirical research applies the criterion of empirical accuracy. Yet, when it is difficult to establish what the relevant goals are or when it is unclear whether behavior is successful when assessed against such goals, it seems that the perspective offered by AR fails to provide practical tools for the assessment of decision-making as and when it happens.[65] Consider, for instance, that it might

[65] Interestingly, Shira Elqayam (Elqayam and Evans 2011; Elqayam 2012) has offered an interesting discussion of some problems of MBR by appealing to themes that are, at least in some respects, similar to those discussed by AR theorists, as she also takes SPR to be not necessarily linked with instrumentally rational behaviour. At the same time, her conclusions also seem to differ from those drawn by AR theorists. Specifically, while AR theorists believe that researchers ought to explore whether or not behaviour and cognition are successful, Elqayam seems to be less inclined to accept the use of normative considerations when exploring people's thinking and decision-making, encouraging instead a merely (or mostly) descriptive approach in the study of judgement and decision-making. It seems to me that Elqayam's scepticism about the possibility of normative assessment is due to the fact that she focuses on a number of contexts in which the assessment of behaviour in light of people's goals is rather difficult, whilst AR theorists have mainly focused on cases in which such assessment was more easily carried out.

prove extremely difficult to say something precise about the effect a particular human behavior has on evolutionary fitness.[66]

It seems to me that this objection is plausible, and that these are legitimate worries. But this does not imply that we should refrain from trying and assessing whether people's behavior is adaptive and successful, even if in some cases this kind of assessment might prove to be particularly difficult. The conclusion that MBR researchers should be careful to analyze the structure of the agent's environment as well as her goals before claiming biases still holds.

5. Conclusion

In this chapter, I have addressed the question of whether biases reported in MBR should be seen as instances of irrational behavior and cognition. More precisely, I sought to defend the AR theorists' contention that behavior that violates the norms of SPR can be successful and adaptive. In light of this, I have argued that evaluating reasoning and decision-making against SPR is often normatively problematic and challenged the view that SPR provides us with adequate normative benchmarks for the study of rational behavior and cognition.[67]

[66] For instance, some of the claims made by AR theorists about the evolutionarily adaptive value of particular traits, such as *risk aversion* (e.g., Hintze et al. 2015), can be challenged and seem to be indeed controversial (see, e.g., the points raised by Schulz 2008).

[67] A remark is in order here. So far, I have treated the methodological argument discussed in Chapter 1 and the normative argument presented in this chapter as independent. However, it can be argued that the abovementioned normative claims have methodological implications as well. Notably, AR theorists have paid particular attention to the goal of empirical accuracy and assessed how well judgements correspond to features in the external environment. This seems to explain, at least in part, their interest in formats with natural frequencies, rather than single probability formats.

Chapter 4: Adaptive rationality and goal-based rationality

In the previous chapter, I focused on some problems arising from the assessment of behaviour and cognition against the norms of SPR advocated by MBR researchers. In this chapter, I further elaborate on this topic by examining how we can best conceptualize the challenge articulated by AR theorists. In so doing, I prompt AR theorists to build their case on more solid conceptual grounds. Firstly, I show that AR theorists often appeal to a distinction between coherence and correspondence criteria of rationality originally introduced by Hammond to explicate their normative challenge. Secondly, I argue that a distinction between rule-based and goal-based rationality might better explicate their normative challenge. Thirdly, I point to some unresolved issues that need to be addressed in future research within the framework of AR.

1. Introduction

As we have seen, according to AR theorists what really matters for the assessment of behavior is whether this is adaptive, and not whether it complies with a set of conventional principles of rationality. Specifically, AR theorists argue against the use of norms of SPR as benchmarks of rationality for the study of behavior and cognition. The contrast between these different approaches on rational behavior and cognition is rarely spelled out in detail, though. In fact, focus on adaptiveness is generically contrasted to focus on conformity to SPR. This trend is quite common in the literature. For instance, Ayton and Fischer write that:

> Studying beliefs by comparison with normative models and studying the adaptiveness of their associated behaviors can lead to dissociable conclusions about the efficacy of cognition. (2004, 1377)

Recently, however, AR theorists have tried to articulate their normative challenge to MBR in more detail. To do so, they have appealed to a distinction between coherence and correspondence criteria of rationality originally introduced by Hammond (1996; 2007). For instance, AR theorist Gigerenzer latches onto this distinction when he writes that:

> We do not compare human judgment with the laws of logic or probability, but rather examine how it fares in real world environments. The function of heuristics is not to be coherent. Rather, their function is to make reasonable, adaptive inferences about the real social and physical world given limited time and knowledge. Hence, we should evaluate the performance of heuristics by criteria that reflect this function. Measures that relate decision-making strategies to the external world are called correspondence criteria (Hammond, 1996). (1999, 22)

The idea here seems to be that what we have dubbed so far SPR represents 'a championing of coherence criteria', as philosopher Rysiew once wrote (2008, 1165). More precisely, it seems that the labels "coherence" and "correspondence" are here supposed to capture something deep about the nature of the standards of rationality that have been used, and of those that ought to be used, in the assessment of human rationality. In this chapter I will try to make clearer to what exactly these labels refer. Moreover, I will also seek to show that, how, and why, the coherence - correspondence distinction fails to provide a useful conceptual framework to explicate the AR challenge to MBR.

The first step, here, is to briefly trace the historical development of the coherence - correspondence distinction since its original introduction. Notably, Hammond (1996; 2007) first introduced the coherence - correspondence distinction to identify two strategies available in the study of human judgment. One is called correspondence 'because it evaluates the correspondence between the judgment and the empirical fact that is the object of the judgment' (2007, XVI). Coherence, on the other hand, has to do with the fit between people's judgments. Specifically, Hammond (2007, XVI) defines coherence as 'the consistency of the elements of the person's judgment'. According to Hammond, 'it is easy to see the difference between a judgment that is directed towards coherence—make it all fit together—and one that is directed toward the correspondence between a judgment and a fact' (XIX).

As it turns out, it is not really clear, in Hammond's work, whether his notion of coherence refers only to the fit between beliefs or, rather, whether the idea of 'making it all fit together' includes the fit between beliefs and behaviour as well. As a result, it seems quite difficult to characterize the notion of coherence in a clear and unambiguous way.

To be fair, Hammond was aware that, while correspondence clearly refers here to empirical accuracy, a criterion we are all familiar with, the characterization of coherence was more problematic. In trying to make sense of his notion of coherence, Here I will try to be as charitable as possible. Since Hammond thought that the distinction between coherence and correspondence could prove useful to explain a whole slew of phenomena and to make sense of different lines of research in the

psychological study of human judgment, it seems that a quite loose and broad understanding of Hammond's "fit together" would be preferable. Specifically, according to a loose understanding of the term, even people's beliefs that fail to line up with their actual behaviour seem to violate coherence. After all, a great deal of social psychology research has claimed biases by virtue of having identified discrepancies between what people say they will do and what they do. On the other pole of the distinction is correspondence, as we have seen, and Hammond thought that Brunswik's research on perception, which I briefly described in Chapter 1, offers a prominent example of research on correspondence. As the reader will recall, Brunswik's work focused entirely on the empirical accuracy of physical and social perception—the correspondence between a judgment and an object.

Recently, however, Hammond's distinction between coherence and correspondence has been used and introduced in a broader context. Specifically, the distinction has been used in the study of both judgement and decision-making, and not only in the study of judgement. More generally, the distinction has been introduced in the "rationality debate" to characterize the two competing perspectives on rational behaviour and cognition that we have discussed in the previous chapter: "coherence" has been taken to signify adherence to norms of SPR and "correspondence" the achievement of adaptive behaviour. More precisely, in this context AR theorists stress that this distinction can prove useful to explicate the challenge that they have mounted against MBR. Following AR theorists, the coherence - correspondence distinction has been quite widely praised as a useful conceptual tool in the study of judgment and decision-making and rationality (Baron

2012; Lee and Zhang 2012; Adam and Reyna 2005; Mandel 2005; Newell 2005; Wallin 2013). Unsurprisingly, given the current popularity of this distinction, this distinction has also been celebrated in 2009 in a special issue of the journal *Judgment and Decision Making*.

2. The many meanings of coherence and correspondence

It is also evident, however, that as soon as Hammond's distinction has been applied and used in such broader context, the terms "coherence" and "correspondence" have been modified in important ways. What I want to stress, here, is that Hammond's original distinction has not been just adopted, but also significantly adapted in the literature. Consider, first, the case of coherence. On Hammond's view, coherence seems to refer to a fit between mental states and between mental states and behavior. Yet, with its introduction in a broader debate, the notion of coherence has been also conceptualized in different ways. For instance, Newell writes that 'human judgment can be evaluated by the degree to which it *coheres* with a formal model, such as Bayes theorem, and by the degree to which it *corresponds* with the properties of the environment' (2005, 11). Here coherence does not refer to the fact that judgments fit well with each other, but rather that judgments cohere with formal principles. In a similar vein, AR theorist Jeffrey Stevens points out that 'human judgment can be evaluated by the degree to which it coheres with a formal model, such as Bayes' theorem' (2008, 291). Notably, also other different explications of the concept of coherence have recently been given. For example, Lee and Zhang write that 'the distinction between the assessment of a process and the assessment of an outcome is

presented as a distinction between coherence and correspondence' (2012, 366). These conceptualizations of the coherence – correspondence distinction can hardly refer to the very same things and point to the identical alleged contrasts. This should show that the concept of coherence has become less and less clear.

One might think that this is just hair-splitting. In fact, however, some phenomena that are taken to be instances of coherence in this broader debate seem to bear very little resemblance to Hammond's original characterization. Overall, it seems that coherence has now become a rather complex and heterogeneous mixture. To appreciate this, consider that AR theorists frame their normative challenge in terms of a replacement of traditional focus on coherence with focus on correspondence, and much of their empirical evidence to support such a conclusion is about the successful performance of non-compensatory heuristics. Notably, while tenets of rationality such as transitivity have to do with coherence in the sense of 'fitting together', compensatoriness does not seem to. In fact, one could even say that non-compensatory strategies could be perfectly coherent as a matter of logical consistency and of coherence of preferences.

To be fair, the situation does not look much better when we consider the case of correspondence. Recall that, on Hammond's view, correspondence referred to empirical accuracy only. Now, some authors have preserved Hammond's characterization. For instance, Stevens (2008, 291) claims quite clearly that 'correspondence refers to the degree to which decisions achieve empirical accuracy; that is, whether they reflect the true state of the world'. But others have used

the concept of correspondence more broadly, stressing that 'there are multiple correspondence criteria relating to real-world decision performance' (Todd and Gigerenzer, 2000, 738), where these encompass empirical accuracy, speed, frugality, fitness maximization, and successful exchanges with others.[68] This is explicit in the work of many AR theorists (though not all of them, as the quote of Stevens reveals), who tend to stress that behaviour and cognition ought to be assessed against correspondence and that correspondence refers to speed, accuracy and frugality among other goals. At other times, AR theorists suggest that correspondence just refers to the achievement of a goal (cf. Gigerenzer and Gaissmeier 2011, 458). It should be clear, therefore, that AR theorists do not typically tend to refer to empirical accuracy only when using the label "correspondence".

Overall, it is evident that both coherence and correspondence have undergone a significant conceptual change. This does not automatically imply that we have to abandon the distinction. Conceptual change is in fact part of scientific progress, one might argue. Moreover, one might claim that these semantic changes do not have to be necessarily seen as an oversight or as resulting from a lack of care in the application of Hammond's original concepts. In fact, one might argue that, to some extent, it should come as no surprise that these notions have been reshaped. For

[68] In characterizing the perspective of AR theorists, scholars such as Samuels, Stich and Bishop (2002) write that 'according to Gigerenzer, a central consideration when evaluating reasoning is its accuracy' (255). Similar considerations can also be found in Rysiew (2008, 1161). In light of these considerations, these scholars have also established links between reliabilism in epistemology and the AR project in psychology. For instance, Samuels, Stich and Bishop write that 'Gigerenzer's accuracy-based criterion for epistemic evaluation bears an intimate relationship to the reliabilist tradition in epistemology' (2001, 255). However, it is important to keep in mind that, while accuracy does seem to be a crucial concern for AR, these scholars are also interested in other goals and standards.

instance, Hammond's original definition of coherence as the internal fit of someone's judgments was admittedly vague and thus allowed for quite flexible use.

However, the problem here is not just that these terms have changed their meaning, but rather that it is unclear what these terms are now even supposed to mean. When AR theorists and other researchers engaged in current debates over rational judgement and decision-making talk about correspondence, it is unclear whether they are referring to empirical accuracy or whether they are pointing to other goods as well. Moreover, if they are referring to other goods, it is still left underspecified what these other goods are. Quite frustratingly, several authors switch back and forth between Hammond's conceptualization of correspondence as empirical accuracy and other and more liberal characterizations.

It should come as no surprise that, given this lack of clarity, it is quite difficult to answer key questions about the nature of rational behaviour and cognition. For instance, is achieving correspondence tantamount to achieving adaptive behaviour? It is not clear how to answer. It seems fair to claim that, if correspondence refers to empirical accuracy, there can be reasons to avoid answering in the positive. To see this, it might be worth recalling some of the considerations expressed in our previous chapters. As we have seen in Chapter 2, in some contexts paying heed to empirical accuracy only might turn out to be evolutionarily maladaptive. But empirical inaccuracy could be not only evolutionarily adaptive. In fact, some scholars have also argued that empirically inaccurate judgement might promote other goals as well, such as mental health (Taylor 1989).

Another implication of this lack of clarity in the characterization of correspondence is that it is unclear whether such criteria of accuracy apply to preferences as well as to inferences. In particular, Stevens points out that 'both criteria (coherence and correspondence) apply to inferences; preferences, however, have no correspondence criteria' (2008, 291). To be sure, if we take correspondence to refer to empirical accuracy, these conclusions seem to follow quite naturally. The underlying argument is the following. You might ask, for instance, whether someone's preferences are transitive or not. But preferences cannot be assessed in terms of empirical accuracy. In brief, only beliefs can be true of the real world: desires and preferences cannot.[69] If AR theorists seek to replace focus on coherence with focus on correspondence, yet correspondence refers to empirical accuracy, there is no room for an appraisal of the rationality of people's preferences.

It is important to stress, however, that if there are multiple and different correspondence criteria—and not empirical accuracy only—it is unclear why correspondence would and should apply to beliefs only. After all, the idea that rationality applies to preferences as well is not new. Dennett, for example, once voiced a concern with traditional approaches to rationality: 'a system's desires are those it ought to have, given its biological needs and the most practical means of satisfying them' (1987, 49). Moreover, as Hausman puts it, 'preferences, like judgements, are subject to rational scrutiny' (2011, 117). Also Frankfurt made the

[69] It is worth noting that in much of the philosophy of mind and of the social sciences, the terms "desire" and "preference" are used interchangeably. However, there could in fact be good reasons to avoid this equation. For instance, preferences are two-place relations, desires only one-place relations. For some critical discussion of the relationship between desire and preference, see Schulz (forthcoming).

point that we ought to have certain desires—although his argument is related to his notion of "caring" (Fankfurt 1982). This matters quite a lot: while it is true that empirical accuracy does not apply preferences, other criteria of correspondence could apply, so to speak, to preferences. This comes out quite clearly, I think, if we consider the criterion of fitness maximization (but other criteria seem to apply too), which seems to apply to preferences. Take, as an illustration, the case of mate preferences. Evolutionary psychologists have investigated them at length (e.g., Buss 1989; Kenrick and Keefe 1992). Mate preferences can affect the current direction of sexual selection by influencing who is differently excluded; they may also reflect prior selection pressures, and exert selective pressures on other components. Overall, it seems that preferences can vary in their degree of adaptiveness.

It should come out quite clearly that, by employing the coherence - correspondence distinction, researchers have not pushed discussions over human rationality forward, and have not promoted clarity in the "rationality debate". Hammond's original concepts of coherence and correspondence are far too narrow tools for the purposes of the AR project, viz. to replace traditional standards of rational behaviour and cognition with criteria of adaptiveness. But other attempts to extend the scope and meaning of the coherence - correspondence distinction have also created confusion in the literature.

There seems to be an interesting consideration to make here. As we have seen, AR theorists' adoption of such an unclear conceptual framework appears to hinder progress in the "rationality debate". This is particularly striking if we consider that

the very same AR theorists who have distorted Hammond's coherence - correspondence distinction and explicated their normative attack on MBR so poorly have also run important campaigns in defence of conceptual and methodological rigor in the study of judgment and decision-making. For instance, AR theorist Gigerenzer argued that 'the heuristics in the *heuristics-and-biases* program are too vague to count as explanations' (1996b, 553). Specifically, he argues that, because Kahneman and Tversky's heuristics are formulated by means of vague terms like representativeness, the appeal to these heuristics as generators of biases has limited explanatory power. Moreover, and more recently, AR theorists have also argued against recent trends to account for people's judgement and decision-making by adopting dual process frameworks. According to Marewski, Gaissmeier and Gigerenzer, 'a dual process framework is too vague to be useful' (2010, 178).[70] Alas, it seems fair to conclude that these scholars have not managed to live up to their high standards of rigor when articulating their normative perspective.

3. In search of a better conceptual framework

Granted that appealing to the coherence - correspondence distinction has not helped to explain the challenge that AR theorists are talking about, the question that arises is

[70] In this work I am not going to discuss the nature of dual process frameworks in detail. Suffice it to say, for our purposes here, that according to dual process theories there are two different kinds of cognitive processes underlying human cognition, and thus two distinct processing modes available for a cognitive task. The first process is characterized by low computational expense and autonomy. The second one is non-autonomous and computationally expensive. Stanovich (1999) famously dubbed these different cognitive processes as System 1 and System 2. According to dual process theorists, cognitive biases should be attributed mainly to System 1 processing, although these scholars have recently started to refine this claim (e.g. Evans and Stanovich 2013). In particular, Stanovich stresses that System 1 was designed for the promotion of narrowly genetic goals, such as reproductive success, whereas the more flexible System 2 serves the goals of the individual person and allows up to rebel against genetic imperatives. For some critical discussion of dual process theories, see, e.g., Carruthers (2012), Osman (2004; 2013), and Kruglanski (2013).

whether there are better ways of explicating it.[71] It seems to me that the best way of interpreting and explicating such challenge is to appeal to a contrast between the following two approaches to the assessment of rational behaviour and cognition. One based on conformity with rules of rationality, and another one based on the achievement of goals and desired outcomes. The best way to conceptualize that challenge is, therefore, to claim that behaviour should be assessed in terms of the achievement of people's goals and desired outcomes, and not against a set of rational norms. I suggest that we refer to these different approaches as rule-based and goal-based rationality. Importantly, this distinction should not suggest that following rules could not result in the achievement of goals. What this should suggest, instead, is that behaviour should be assessed against goals rather than rules.[72]

Interestingly, contrasts analogous to the one between what I am calling here rule-based and goal-based rationality seem to underlie several other distinctions that have been offered in the literature. For instance, Oaksford and Chater, the advocates of the

[71] Note, however, that here I am not claiming that the distinction between coherence and correspondence cannot be useful in other debates. For instance, the concepts of "coherence" and "correspondence" are employed in epistemology (e.g., Ollson 2005), although there is not necessarily one-to-one mapping between Hammond's use of these concepts and how they are employed in those debates.

[72] This does not mean that there cannot be implicit goals in following rule-based rationality. As Gigerenzer himself points out: 'each of the […] systems pictures the goals of human behavior in its own way. Logic focuses on truth preservation. Consequently, mental logic, mental models (in the sense of Johnson-Laird, 1983), and other logic-inspired systems investigate cognition in terms of its ability to solve syllogisms, maintain consistency between beliefs, and follow truth table logic. […] Probability theory depicts the mind as solving a broader set of goals, performing inductive rather than deductive inference, dealing with samples of information involving error rather than full information that is error-free, and making risky "bets" on the world rather than deducing true consequences from assumptions' (2008, 20). Yet, it is worth noting that researchers appealing to rule-based rationality typically treat goals such as "consistency" and "truth-preservation" as the relevant goals in all domains and contexts, without making sure that these are the goals (or the only goals) entertained by the reasoners. Scholars appealing to goal-based rationality, on the other hand, tend to appeal to a broader range of goals, and such goals are not taken to apply universally, since the importance of a goal seems to be context-dependent.

"rational analysis" project presented in our previous chapter, claim that according to standard perspectives on rational behaviour:

> To be rational is to reason according to rules (Brown, 1989). Logic and mathematics provide the normative rules that tell us how we should reason. Rationality therefore seems to demand that the human cognitive system embodies the rules of logic and mathematics. (1994, 608)

This quote seems to provide a characterization of something similar to what I have been calling here rule-based rationality. Notably, at other times the very same authors refer to "formal rationality" to signify the perspective on rational behaviour characterized by conformity with rational norms. According to them:

> If formal rationality is viewed as basic, then the degree to which people behave rationally can be assessed by comparing performance against the canons of the relevant normative theory. (Chater and Oaksford 2000, 99)

But a contrast between rule-based and goal-based rationality seems to underlie several other distinctions that had been offered in the literature. For instance, Evans and Over (1996) famously distinguished between personal rationality (rationality 1) and impersonal rationality (rationality 2). The authors characterize the former as thinking or 'acting in a way that is generally efficient for achieving ones goals', and the latter as 'thinking or acting when sanctioned by a normative theory' (1996, 8). Moreover, Samuels, Stich and Faucher (2004) appealed instead to a distinction between deontological and consequentialist approaches to rational behaviour. According to the deontologist, what it is to reason correctly—what is constitutive of good reasoning—is to reason in accord with some appropriate set of rules or principles, while another prominent view, which is often called consequentialism,

maintains that what it is to reason correctly is to reason in such a way that you are likely to attain certain goals or outcomes.[73] In addition, Kacelnik, Schuck-Pain and Pompilio (2006), in the context of evolutionary biology, have contrasted conformity with the consistency axioms of economics with the achievement of organisms' adaptive goals.[74] While these distinctions might present some differences, they all seem to revolve around a contrast between two different questions: Is the organism following the relevant norms of rationality? Is the organism achieving his goal? While researchers interested in exploring, say, the *conjunction fallacy*, seem the be interested in the former question, AR theorists seem to be interested in addressing the second kind of question when asking whether people are making quick and accurate predictions in the real world.[75]

Now, what I want to suggest is that such a contrast between rule-based and goal-based rationality could be used to frame the challenge mounted by AR theorists as well. I also contend here, however, that this characterization should be seen only as a starting point, as it needs further refinement.

---

[73] It is worth mentioning that the distinction between deontological and consequentialist has been applied in the context of the "rationality debate" by Schurz (2014) as well.

[74] Although all these distinctions are binary, these contrasts should not be necessarily seen as exhaustive. For instance, one could assess behavior by focusing on intellectual virtues such as open-mindedness and attentiveness, which are not clearly goals entertained by people or formal rules.

[75] One worry one might have here is that the definition of "instrumental rationality" I offered in the Introduction is actually very close to what I am calling here goal-based rationality. Some remarks are in order, though. Notably, here I am using the labels of rule-based and goal-based rationality to move beyond the coherence - correspondence distinction, and my suggested distinction seems to be rather helpful in capturing what the genuine bones of contention in this debate are: should we assess behaviour against formal rules or goals? Moreover, in my presentation of goal-based rationality I am pointing towards an attitude and stance to be adopted in the assessment of behaviour and cognition: behaviour has to be assessed against people's goals. This methodological aspect was not part of the characterization of "instrumental rationality".

First, an obvious worry to have with regard to rule-based rationality is that it might be unclear what logic, probability theory, and expected utility theory we are talking about. For instance, if your logic is of some paraconsistent variety, then a subset of your beliefs may well be inconsistent. Moreover, consider that some theoreticians (e.g., Fishburn 1991; Bordley and Hazen 1991; Anand 1987) claim that it could be rational to violate transitivity. In addition, compensatoriness does not seem to be necessarily part of rule-based rationality. This is important, as the most robust evidence invoked by AR theorists comes from research on non-compensatory strategies that can be fast, frugal and accurate. The objection might seem to be an important one. Yet, I would like to make a couple of remarks. First, while it is true that the characterization cries out for further clarification, it seems that there is some general consensus around a set of principles of rationality. For instance, consider the *conjunction rule* in probability theory: the probability of some event A occurring cannot be less than the probability of some other event and A both occurring. Moreover, consider the *principle of descriptive invariance*, which states that the preference order between prospects should not depend on the manner in which they are described. Second, even if there was disagreement on what norms should count as a rational principles and be part of rule-based rationality, this would not be particularly damaging to my analysis here: what I am trying to claim is just that, according to a popular perspective on rational behaviour and cognition, the latter has to be assessed against a set of rational norms.

Another worry one might have here is that it is unclear whether rule-based rationality requires people to actually use a rule in their conscious reasoning. This

does not seem to be the case. If we are looking here for a general characterization of the approach applied in research on rational behaviour, it seems that researchers have not required people to use norms in their conscious reasoning. In fact, it is worth noting that rule-based rationality has been applied quite broadly also to ranges of animals and even in cases in which it is unclear to what extent they might display conscious reasoning (cf. Stanovich 2013; Alcock 2005; Arkes and Ayton, 1999; Dukas, 1998; Fantino and Stolarz-Fantino, 2005; Hurley and Nudds, 2006).[76]

Moreover, since AR theorists seek to offer the perspective of goal-based rationality as an alternative to traditional standards of rationality, or what I have called rule-based rationality, this perspective has to be characterized quite in detail. The reader should recall here that, according to this view, behaviour and cognition should be assessed in terms of the achievement of people's goals. As I have argued in previous sections of this work, ideas similar to the ones expressed by AR theorists have recently started to gain popularity. In particular, Elqayam and Over write, along similar lines, that if 'behavior is typically well adapted and people typically achieve their personal goals, it can be described in some terms as rational' (2011, 236-7). An appeal to the importance of goals has characterized the project of "rational analysis" too: Chater and Oaksford (2000, 93) present their empirical program as 'attempting to explain why the cognitive system is adaptive, with respect to its goals', although they suggest that people's success at achieving goals is ultimately due to their

[76] There is an interesting open debate over which animals have consciousness and what (if anything) that consciousness might be like (cf. Lurz 2009).

following standard rational principles. Thus, it seems that AR theorists are not alone in appealing to the relevant notion of "goal".[77]

Alas, it also turns out that the characterizations of what goals are which are generally offered in the literature are quite unsatisfying. As a result, also the perspective of goal-based rationality still seems to cry out for further clarification and refinement. Very few theoretical treatments of the notion of goal have been offered in the psychological literature. With regard to this point, Castelfranchi laments that:

> A theory of goals be a prerequisite for a principled theory of decision-making might look obvious and already well acknowledged in the literature (van Osselaer et al. 2005; Kruglanski 1996), but it is not so. The state of our ontology and theory of goals for an adequate description of decision-making is really disappointing: a lot of distinctions and of clear process models are still needed. (2014, 104)

As this quote seems to reveal, the lack of a theory of what goals are seems to affect most psychological research on decision-making that appeals to such construct, and not just the AR project. This does not mean, however, that no tentative accounts have been offered. For instance, a notable attempt to offer a characterization is due to Kruglanski and Kopetz (2009, 28), who point out that:

[77] Other researchers have recently emphasized the importance of assessing behaviour and cognition against goal-based rationality, and a number of authors have suggested a list of outcomes that are supposed to be suggestive of adaptive and successful behaviour. In particular, according to authors in the "Fundamental Motives Framework" (Kenrick et al. 2010), humans have inherited psychological mechanisms for solving a set of specific ancestral social challenges and achieve their goals. These fundamental challenges include: (1) evading physical harm, (2) avoiding disease, (3) making friends, (4) attaining status, (5) acquiring a mate, (6) keeping that mate, and (7) caring for family. While this perspective is motivated by evolutionary considerations, goals such as avoiding diseases seem to be highly relevant to people's instrumental rationality as well.

> We define "goals" as subjectively desirable states of affairs that the individual intends to attain through action (Kruglanski, 1996). The notion that the individual intends to attain a goal implies that the goal is perceived as attainable, apart from it being perceived as desirable. In other words, adoption of a goal entails a process of proof (Kruglanski, Pierro, Mannetti, Erb, and Chun, 2007) in which the individual confronts subjectively relevant evidence as to a given state's desirability and attainability. (2009, 28)

Notably, even if one accepted this characterization, there would still be several questions one may want to ask. For example, one question that arises, and that the abovementioned definition does not explicitly address, is whether conscious decision-making is required for the individual to attain a goal. It is true that most theories of goals emphasize the role of conscious choice in the adoption of goals and that goal adoption needs to be accompanied by a conscious decision. However, according to some recent accounts, goals can be automatically put in place by situational cues, and they can guide behaviour without a person's awareness of the operation of these goals (Aarts and Dijksterhuis, 2000; Bargh, 1997; Bargh and Gollwitzer, 1994).

But there is an even more important issue lurking here. While the abovementioned researchers disagree on whether conscious decision-making is required for the individual in order to attain her goals, these researchers share the view that, whether conscious or not, goals are still psychological states. Yet, different understandings seem available.[78] In particular, according to other scholars achieving a goal means just doing well on some external performance measure. Notably, whilst on some occasions AR theorists have seemed to switch between these two different

[78] For some recent discussions of different conceptualizations of goal-directed behaviour, please see Butterfill and Apperly, (2013, 613-614).

understandings, in recent contributions it has been stated quite clearly that goals should be conceptualized as the fulfilment of some psychological desire-like state (e.g., Gigerenzer and Sturm 2012). Here I just want to draw the reader's attention to an often-overlooked implication of such a characterization. Specifically, it seems that a characterization of goal-based rationality that conceives of goals as psychological desire-like states would be too demanding to be applied to several non-human animals: AR theorists have often tried to apply the normative perspective of goal-based rationality to non-human animals as well,[79] but in many cases concepts such as desires or intentions might not be appropriate or applicable. Other moves may be available to AR theorists to overcome this impasse. For instance, AR theorists might want to offer a bifurcation of what we have called goal-based rationality. When this applies to human animals, goals are meant to refer to the fulfilment of some desire-like state. On the other hand, when it refers to at least some non-human animals, goals refer to some external performance measure related to survival and reproduction.[80]

To be sure, these are not trivial issues, and providing a clearer characterization of the construct of "goal" and of the relevant notion of "goal attainment" seems to be an

[79] Given these researchers' interest in instrumentally rational behaviour, this choice seems, at least in some cases, quite plausible. After all, a number of animals in some cases do seem to exhibit instrumentally rational behaviour. For instance, Searle (2001) begins a book on the philosophy of rationality by referring to the famous chimpanzees on the island of Tenerife studied by Wolfgang Kohler (1927). Several of the feats of problem solving that the chimps displayed have become classics and are often discussed in psychology textbooks. In one situation a chimp was presented with a box, a stick, and a bunch of bananas high out of reach. The chimp figured out that he should position the box under the bananas, climb up on it, and use the stick to bring down the bananas. Searle (2001) asks us to appreciate how the chimp's behaviour fulfilled all of the criteria of instrumental rationality—the chimp used efficient means to achieve its ends. The "desire" of obtaining the bananas was satisfied by taking the appropriate action.

[80] After all, the idea that animals exhibit rationality in different degrees is not new (cf. Bermudez 2003, chap. 6).

important task for the AR project. Acknowledging that more research is needed should not mean, however, that researchers should refrain from trying to assess behaviour and cognition by focusing on the achievement of goals and desirable outcomes. Experimental work on the assessment of judgement and decision-making and theoretical work on what counts as a goal and as goal-attainment should proceed in parallel.

4. Conclusion

In this chapter I have focused on how we can best explicate the normative challenge to MBR that has been put forth by AR theorists. I have argued that, while AR theorists generally appeal to a distinction between coherence and correspondence criteria to explicate the challenge they are talking about, there are reasons to avoid using such distinction. In fact, the introduction of this distinction in the debate has resulted in confusion and hindered progress in the field. I have instead suggested that researchers should appeal to a distinction between rule-based and goal-based rationality and identify unsuccessful behaviour with failures of goal-based rationality. I have also shown, however, that much more work still has to be done in order to further clarify the relevant notions of rule-based and goal-based rationality.

Chapter 5: Adaptive rationality and cognitive limitations

I have suggested that, in a number of domains, behavior and cognition seem to be adaptive in spite of the fact that they violate what I have dubbed rule-based rationality, and that researchers interested in adaptive behavior should try to assess performance against people's goals. However, one might argue that my characterization of the AR project has gone amiss, and that the *crux* of the debate over the plausibility of AR lies elsewhere. Or one might simply argue that I have at least ignored a key argument for AR here. What specifically would I have missed here? The objector might argue here that the main problem with rule-based rationality is that there are important limitations hindering rule-based rationality from being reasonably applied. The goal of this chapter is to deflect these potential criticisms. I do so by first exploring the relationship between literature on so-called "bounded rationality" (henceforth, BR) and AR. Specifically, in what follows I start by showing that commentators in the "rationality debate" often appeal to versions of the *ought-implies-can* principle to defend the normative significance of research on BR, and on occasion interpret the normative challenge mounted by AR theorists by reference to such principle. I then argue that such an appeal is not needed. In fact, the normative framework of AR offers a straightforward way of interpreting the normative significance of literature on BR that does not appeal to versions of the *ought-implies-can* principle.

1. Introduction

In outlining the AR challenge to MBR, I have not as yet appealed to human cognitive limitations. But one might wonder, at this point, whether my characterization has missed an important part of the debate—perhaps even its very core. After all, some authors refer to AR theorists by employing the very label "bounded rationality" (BR) (e.g., Sturm 2012). To be sure, AR theorists often offer considerations about the bounded nature of human cognition. For instance, Gigerenzer and Goldstein write that 'unbounded rationality is a strange and demanding beast' and that 'cognitive algorithms need to meet more important constraints than internal consistency: they need to be psychologically plausible' (1996, 665). Moreover, AR theorist Gigerenzer stresses that 'Bayes' rule and other rational algorithms quickly become complex and cognitively intractable, at least for ordinary human minds' (1996, 277). Unsurprisingly, in light of these and similar points, the AR project has often been seen to rely on the importance of cognitive limitations. For instance, Newell claims that, according to AR theorists, 'the methods of classical rationality are computationally intractable and time consuming, and thus beyond the bounds of human decision makers' (2005, 11).

The reader might find herself rather confused: is it not the very fact that people have cognitive limitations and that their reasoning is bounded already an admission that they cannot be rational? After all, MBR researchers do seem to accept that people's reasoning is bounded, but from this premise they draw rather pessimistic conclusions about human rationality. For instance, Richard Thaler explains that 'Kahneman and Tversky have shown that mental illusions should be considered the

rule rather than the exception. Systematic, predictable differences between normative models of behaviour and actual behaviour occur because of what Herbert Simon called bounded rationality' (1980, 40).[81]

According to a number of researchers engaged in the "rationality debate", however, we should refrain from drawing pessimistic conclusions from such considerations. Specifically, many scholars have emphasized that normative standards need to take into account the resources of real cognizers, and that departures from normative standards that are unreasonably demanding cannot or should not count as cases of irrational behavior. For instance, Nozick, when commenting on the traditional normative perspective on rationality, writes that 'it seems necessary that it be a theory which can be satisfied by someone: that is, that it not be a theory which is such that in order to satisfy it a being would have to possess powers, capacities and skills far beyond those possessed by humans beings as they now are' (1963, 24). Moreover, Cherniak (1986) has argued that the conception of rationality that has been pervasively assumed is too idealized, and that only "minimal rationality" is required for an actual theory of cognition. In addition, Good (1993) has made allowances for human limitations in his distinction between type 1 rationality, which demands complete conformity with the axioms of utility theory or probability, and type 2 rationality, which takes into account the real cost of theorising and requires only that no contradiction is found. But similar concerns seem to motivate also the distinction that is sometimes made between normative and

[81] Other scholars have pointed instead towards a tension between the very concepts of boundedness and rationality. In particular, in the words of Morton: 'Do not the boundedness and the rationality pull in opposite directions, boundedness lowering standards and rationality raising them' (2012, 176)?

prescriptive models of rationality (e.g., Baron 1985; Stanovich 1999). While normative models show how the thinker would do in specific situations without taking into account human limitations and constraints, prescriptive models show how the ideal thinker should reason once human limitations and constraints are taken into account. Yet the list of authors who have made similar points is rather large, including Simon (1957; 1973; 1983; 1990) and Harman (1986; 1995). The general idea that characterizes these otherwise heterogeneous sets of perspectives on rational behaviour and cognition is that we should not ignore the fact that, whilst God and Laplace's super-intelligence do not have to worry about limited time, knowledge, and computational capacities, human beings do. Human rationality is "bounded", and we have to deal with what Cherniak (1986) has referred to as the "finitary predicament" of having limited cognitive resources and time.

In the remainder of this chapter, I will elucidate the nature of the relationship between BR and AR. In particular, after having introduced research on BR, I will present two distinct ways of interpreting concerns about the demandingness of rule-based rationality. I will then show how (and why) the framework of AR outlined so far does not need to appeal to versions of the so-called *ought-implies-can* principle to vindicate the importance of research on BR.

2. The bounds of rationality

The volume of research that goes under the heading of BR is impressive and still growing. Consider, for instance, that a recent special issue of the *Journal of*

Economic Methodology was devoted to the topic *Methodologies of Bounded Rationality*.[82] In this section I introduce the concept of BR and the literature surrounding it, showing that—and how—research on BR is rather heterogeneous and encompasses several different projects. Unsurprisingly, a number of scholars have charged BR research with being unclear and overly vague. For instance, Watts pointed out that 'there are so many ways in which rationality can be bounded that we can never be sure we have the right one' (2003, 66). In later sections I will show that normative interpretations of research on BR focus on very peculiar aspects of descriptive research on BR.

Since the label BR is closely tied to the work carried out by Simon, it makes sense to start our discussion here by looking at his work. Following the presentation of Grüne Yanoff (2007, 543), it is possible to claim that, when Simon introduced the notion of BR, he referred to the fact that humans, unlike fictional omniscient Laplacian demons, have to worry about:

- Limited knowledge of the world;
- Limited time available;
- Limited ability to evoke this knowledge;
- Limited ability to work out consequences of actions;
- Limited ability to conjure up possible courses of action;
- Limited ability to cope with uncertainty;
- Limited ability to adjudicate among competing wants.

[82] I am referring to Volume 21, Issue 4 of 2014.

The core of Simon's proposal was that, besides being physically and biologically limited, human beings are also cognitively limited beings: not only can we not be in two places at once or live forever, we can also only closely pay attention to one thing at a time, and our working memory can only hold the famous 7 ± 2 chunks (Miller 1956). Simon (1955) himself used the analogy of a physical constraint: consider a bird that can only fly at up to 70 kilometres per hour; the set of alternatives available to escape from a predator does not include flight at 200 kilometres per hour. Analogously, a decision-maker cannot compare more than a certain number of alternatives. She must define a boundary for the comparative process and then choose within this boundary. Human rationality is constrained, and hence "bounded", since we possess limited computational ability and selective memory and perception.

However, other lines of empirical research have been associated with the concept of BR. For instance, another dimension of BR was presented in Chapter 1. According to research within AR, a problem can be efficiently solved if represented one way but not if represented another way. Recall, in particular, Gigerenzer's discussion of the difference between single event probabilities and natural frequencies. Let us remind ourselves of the nature of this effect. Suppose your insurance company asks you to take an HIV test, and that this turns out to be positive. You go and visit your doctor, who tells you that there is a 0.01% chance that this result is a false positive and that you are not infected. Moreover, she also informs you that the prevalence of such infection within the group to which you

belong is 0.01%. Would you feel comforted or not? People tend to not be comforted at all. The relevant data about base rates are seemingly ignored by people. However, according to AR theorists, things change rather a lot once such information is provided in terms of natural frequencies. One woman in 10,000 is supposed to have HIV, and her test will be positive. Of the remaining 9,999 women who have the test but do not have HIV, one will also get a positive result. Once this information is provided, it seems that people fare better at figuring out that, in spite of your positive result, there is a 1 in 2 chance that you do not have HIV. What is important to stress here is that, according to this line of research, our cognition is bounded by an alleged inability to deal with some specific problem representations.

It is also worth noting, however, that a number of scholars stress that there are obvious bounds on rationality that do not derive from limited thinking power. We are all subject to wishful thinking, for example. Recently, for instance, the idea that our cognition is bounded was introduced into the moral domain as well, where the term of "bounded ethicality" was coined (Bazerman and Banaji, 2004; Chugh, Bazerman and Banaji, 2005). This is relevant because of the emphasis this notion places on motivational factors. In the words of Chugh, 'bounded ethicality represents a subset of bounded rationality situations in which the self is central and therefore, motivation is most likely to play a prominent role' (2005, 8). More precisely, the authors stress that they:

> Favor a particular vision of the self in our judgments. Just as the *heuristics-and-biases* tradition took bounded rationality and specified a set of systematic,

> cognitive deviations from full rationality, we endeavor to take bounded ethicality and specific systematic, motivational deviations from full ethicality. (...) Ethical decisions are biased by a stubborn view of oneself as moral, competent, and deserving […] To the self, a view of morality ensures that the decision-maker resists temptations for unfair gain; a view of competence ensures that the decision-maker qualifies for the role at hand; and a view of deservingness ensures that one's advantages arise from one's merits. (2005, 9-10)

More generally, while traditional research has solely focused on cognitive capacities as sources of BR, recent work on the role of affective and emotional processes has been included under the same heading of BR (Hanoch 2002). In particular, some authors have begun to emphasise the numerous effects of emotions on behaviour. The idea that our rationality is bounded has thus been used to account for heuristics such as the *affect heuristic* described by Finucane et al. (2000).

The concept of BR has recently been associated with limitations of our willpower as well. As Jolls et al. point out, in addition to BR, 'people often display bounded willpower' (1998, 1479), where "bounded willpower" refers to the fact that human beings often take actions that they know to be in conflict with their own long-term interests. Consider research on so-called "ego depletion". What has been suggested is that self-control is alike to a muscle (e.g. Baumeister et al. 2007): just as muscles demand strength and energy in order to exert force for a period of time, acts that demand high self-control also demand strength and energy to be performed. Similarly, as muscles become tired after a period of continued exertion and have limited ability to employ further force, self-control can also become depleted when demands are made of its resources over a period of time.

I have given a rather general overview of research inspired by, or associated with, BR. One might wonder, given that the notion of BR is used to refer to so many different things and projects, whether the concept of BR might still be useful. I will leave these issues aside. I merely point out that, when attempting to draw normative conclusions from claims about the bounded nature of our rationality, we have to be careful and specify which dimension of BR we are referring to.

3. Taking the bounds of cognition (too) seriously

So far, we have shown that there are in fact several lines of empirical research that appeal to the concept of BR, some emphasizing the role of computational limitations, others appealing, for example, to the impact of motivational factors. It is important to note, however, that much of the interest in BR concerns the alleged normative implications of such research. Specifically, in light of the bounded nature of human rationality, a number of researchers have stressed that rule-based rationality is too demanding, as it asks for more than our capacities allow.

But how exactly should we think of the demandingness of rule-based rationality? One common way of understanding the normative significance of the existence of severe bounds on our rationality is by appealing to the some version of the *ought-implies-can* principle. Interestingly, the principle has often been invoked in the "rationality debate". A number of scholars appeal to some version of it when trying to challenge the idea that violating rule-based rationality is necessarily irrational. For instance, Elqayam (2012) writes that:

> Simon's (1957, 1982, 1983) model of bounded rationality refers mainly to the way that human rationality is constrained by what Cherniak (1986) calls 'the finitary predicament'; that is, physical and cognitive limitations on processing. The fact that humans do not live forever, that our brain capacity is finite, dictates that only tractable computations could count as rational. The underlying rationale is the age old "*ought-implies-can*" attributed to Kant (1932/1787). (2011, 402)

Moreover, Stich has colloquially pointed out that 'it seems simply perverse to judge that subjects are doing a bad job of reasoning because they are not using a strategy that requires a brain of the size of a blimp' (1990, 27). In addition, appealing to similar considerations, Stanovich (1999) introduces—in his contribution to the "rationality debate"—a position he refers to as "Apologism", which seems to subscribe to the *ought-implies-can* principle. More precisely, according to the Apologist:

> It seems perverse to call an action irrational when it falls short of optimality because the human brain lacks the computational resources to compute the most efficient response. Ascriptions of irrationality seem appropriate only when it was possible for the person to have done better. (2004, 157)

The principle was first used in ethics but has since been applied quite broadly. It has also attracted a number of criticisms (see Stern 2004; Vranas 2007; Graham 2011 for some critical discussions of the principle). For our current purposes we simply need to stress that the relevance of the principle clearly depends on what counts as falling within a person's capacities and on what counts as exceedingly demanding. Following Till Grüne Yanoff (2007), we can argue that normative standards might be too demanding in the sense that they require us to do things that we find very hard to do, things that constitute significant sacrifices. But normative standards might also

be demanding also in the sense that they require us to do things that we literally cannot do, things that go beyond our capacities. Using Grüne Yanoff's analogy with morality, in relation to which the principle was first applied, it might be argued that morality is too demanding because it requires people to devote the majority of their income to charity, or because it requires someone to save a person whom it is impossible to save. In a similar fashion, norms of rationality might be too demanding because following them is laborious or because we literally cannot follow them.

Notably, the strength of the principle seems to depend on which notion of demandingness we pick. As Grüne Yanoff points, 'uttering "I can't do this" when faced with the annual tax report, for example, does not usually mean that one does not have the capacity to do it and thus should not do it' (2007, 555). In a similar way, while following rule-based rationality might be demanding, in the weaker sense of the term, the person who claims this should expect to be told that this is just the nature of rationality, and that we just have to accept its demandingness. It seems fair to say that many colloquial uses of 'can' do not make the principle compelling in this context. However, the second and stronger sense of demandingness seems to lend more robust grounding to the *ought-implies-can* principle: it is more difficult to accept that someone is irrational when she or he literally could not have done otherwise. In fact, as we noted earlier in the chapter, the application of the *ought-implies-can* principle in the context of the "rationality debate" seems to be based on the stronger sense of demandingness and on a particular aspect and dimension of

research on BR. Specifically, the clearest cases of strong demandingness refer to our computational limitations.

In the remainder of this section I will discuss this strategy in more detail and also mention some of the problems it seems to face. After this, I will present an alternative way of thinking about the normative relevance of research on BR.

To appreciate the apparent plausibility of the strategy, it is worth stressing that there are cases where concerns about the psychological implausibility of rule-based rationality look quite compelling. Consider the *consistency preservation* principle,[83] which seems to be a tenet of rule-based rationality. We can easily see how some tasks in which this principle is involved might quickly become intractable. The computational problem will emerge once a person begins to monitor and to check the consistency of her beliefs. For instance, consider a person with n beliefs. In order to check the pair-wise consistency of n beliefs, n(n-1)/2 pairs have to be compared. But is consistency computationally possible? Assuming a person has 150 stable relationships and holds 20 beliefs about each partner, she holds a total of total of 3,000 beliefs. Consequently, she has to check 4,498,500 pairs concerning their consistency. Assuming that checking a single pair takes one second, she would have to spend 1,562 workdays (8 hours per day) checking the consistency of her beliefs. If we assume that a person has many more beliefs about her close social network (family, friends and so forth), say 100 times as many, and assume this to consist of

[83] Stein characterizes the principle along these lines: 'Suppose we have the intuition that before a person commits herself to some belief *p*, she should check to make sure that *p* is logically compatible with all her other beliefs' (1996, 163).

20 people, she has to check a total of 915,898,600 pairs for consistency. This would amount to 31,802 years spending eight hours every day on checking the consistency of one's beliefs. Since the number of pairs increases exponentially with the number of elements, checking consistency quickly becomes a computationally intractable problem when n increases. Arguably, for mere mortals, being fully consistent in environments involving larger ns becomes computationally impossible. It thus seems that, at least in this case, it is simply impossible, literally impossible, to conform to rule-based rationality.

It is also interesting to stress, though, that while commentators on the "rationality debate" have often interpreted the normative relevance of research on BR by appealing to this application of the *ought-implies-can* principle, this choice might seem to be problematic.

In particular, there are reasons to think that a strategy appealing to cases of strong demandingness cannot really apply to MBR and, in particular, to well-known tasks in the *heuristics-and-biases* literature. To see this, consider that, while a person trying to follow the *consistency preservation* principle would never have enough time to acquire new beliefs, it does not seem to be the case that following rules such as, for instance, the *conjunction rule* would lead to computational explosion. On the contrary, as Stein has convincingly argued (cf. 1996, 248), the *conjunction rule* that we have discussed extensively on several occasions does seem to be a reasonable candidate for the implementation in human brains in real time. On the face of it, it

does not seem to be true that the norms used in such tasks are not 'psychologically plausible'.

Moreover, the idea that it is not exceedingly demanding to comply with rule-based rationality gains support from further considerations that will be explored in the next chapter, and more precisely, from the evidence offered by Keith Stanovich and his co-workers showing that not everyone violates rule-based rationality in MBR tasks. This suggests that it is computationally possible for humans to do so. Specifically, at least some people seem to be able to follow those norms, and thus following rule-based rationality in standard MBR tasks does not seem to go beyond human cognitive capacities, or at least beyond the cognitive capacity of some of the reasoners. It seems fair to conclude, therefore, that the attack on MBR based on the *ought-implies-can* principle and on a strong understanding of demandingness seems to suffer from a serious shortcoming.

One might reply at this point that following those norms goes beyond the cognitive capacities of *some* people, but not of others. More precisely, one might insist that people who fail to comply with rule-based rationality do so because of the computational limitations they face, which differ from those of people who do not violate rule-based rationality. After all, human beings differ greatly with respect to the limitations they face—the limitations of children differ from those of adults, and the limitations of highly intelligent people differ from those whose intelligence is not so high. However, I see some problems with this view.

One problem might be the following: as we will see in the next chapter, for some biases there is no direct correlation between cognitive capacity and susceptibility to bias, where standard tests of intelligence are often taken to provide an operationalization of cognitive and computational capacity. Thus, it might be difficult to explain differences in performance on the part of reasoners in terms of differences in cognitive capacities. In the words of Stanovich, this strategy:

> [w]ill not work for all of the irrational tendencies that have been uncovered in *heuristics-and-biases* literature. This is because some of those biases are not very strongly related with measures of intelligence (2011b, 357)

Moreover, and more importantly, it is also unclear whether people with limited computational power really cannot do any better and bring themselves to follow rule-based rationality. After all, cognitive limitations do not seem to be absolute and the cognitive load that different tasks place on our brains can be attenuated by using or by altering external factors. For instance, although I do not know of any studies on how much these reduce the computational load, it is evident that when we are coupled with external resources our computational powers can increase in significant ways.[84] As it turns out, we tend to deposit the contents of our working memory in the environment. We do that not solely for the purpose of storing information. We deposit such contents in a form upon which we are still able of executing computations. We can do multiplication in our heads with numbers up to 10.

[84] As it turns out, while it is true that research on BR has inspired some applications of the *ought-implies-can* principle, other research within the framework of BR seems to make the application of this principle rather difficult. Specifically, on the one hand, scholars appealing to the *ought-implies-can* principle suggest that it is impossible to comply with rule-based rationality. On the other hand, research on de-biasing inspired by BR suggests that, at least on a number of occasions, it is possible for people to follow rule-based rationality: since *ought* implies *can,* but *can* seems to be possible for humans, there does not seem to be any argument against the normative validity of rule-based rationality.

However, we are unable to go beyond that—and this in spite of the fact that we have all learnt a method that transforms the larger problem into a number of smaller problems, each of which we are capable of solving in our head (e.g., to multiply 34 by 78, first you multiply 8 by 4, then you multiply 8 by 3, etc.). Each of these sub-problems we can easily solve in our heads, but the reason we cannot solve the problem as a whole is that the solution to these four sub-problems must be kept in working memory in order to resolve the final sub-problem, which involves addition of the four products. Moreover, as AR theorists have tried to show, at least in some cases (e.g., in the case with the *base rate fallacy*) it seems possible to make a problem easier to compute by modifying the format of information (e.g., by using natural frequencies instead of single probabilities).[85] If this is correct, it follows that a person's computational power can be improved in some cases. This suggests, in turn, that it is unclear whether, when the right resources and conditions apply, people with limited computational power really cannot conform to rule-based rationality.

4. AR and the bounds of cognition

So far, we have seen that there is a popular way of interpreting the normative significance of research on BR that appeals to versions of the *ought-implies-can* principle. We have also seen, however, that attacks on MBR based on the application of this principle do not seem to be particularly scathing.

[85] We have discussed some of the relevant literature in Chapter 1, but see also Goldstein and Rothschild (2014) for more recent research on bias mitigation and de-biasing mediated by environmental factors.

Here I wish to clarify that, whilst it is also rather common to interpret the challenge mounted by AR theorists as resting on the application of the *ought-implies-can* principle, this attribution does not seem to be appropriate. Hands writes, for instance, that 'ecological rationality[86] did seem to win on *ought-implies-can* grounds' (2014, 13).[87] This is a rather popular trend, as if there were no other possible ways to think of the normative value of research on BR than by appealing to this principle.

I wish to suggest, now, that there is actually no need to apply this principle in order to make sense of the normative relevance of research on BR and of AR theorists' interest in people's BR. In fact, it seems to me that the framework of AR articulated in the previous chapters might offer a different way to do this.

Recall that, from the perspective of AR, what matters is the achievement of adaptive behaviour and cognition, and that people's performance has to be measured against their goals and desired life outcomes. According to this framework, even if rule-based rationality is not overly demanding in the strong sense required by an application of the *ought-implies-can* principle, it still matters rather a lot whether following the norms of rule-based rationality is laborious and demanding. Specifically, we agents should be sensitive to the fact that following some norms is laborious and costs time, energy, and so forth, because (and as long as) these are goods that people value. In other words, if rule-based rationality requires extensive

[86] It is still worth restating that, as I stated in the Introduction, I stick to the label AR, although this perspective is often referred to as "ecological rationality".
[87] See also Rini (2015, 157) and Carruthers (2006, 229) for similar attributions of this principle to AR.

computations, losses of time and huge efforts, it can hardly be seen to be linked to adaptive behaviour and cognition.

To see this, let us follow the presentation offered by Schulz (2011a, 1277), who has stressed that, from the perspective of AR, the success of an agent's behaviour is determined by an equation like the following: S = aB + cD + eF, where B, D and F are the values of a set of relevant goods, and a, c, and e their relative importance. According to AR, rule-based rationality cannot be conducive to adaptive behaviour and cognition if it overlooks key factors and goals, such as saving time and effort. On this view, reasoning rules should thus be 'psychologically plausible' or, more precisely, they should not be overly demanding. Otherwise, following rule-based rationality would end up being maladaptive.

BR has normative relevance, on this view, because cognitive and time limitations are part of our epistemic context, and the quality of decision-making strategies should be assessed within appropriate epistemic contexts. Given the existence of such limitations, behaviour that violates rule-based rationality can be more adaptive than behaviour that conforms to it.

These aspects seem to be particularly clear in the studies by Herbert Simon. His work on satisficing clearly illustrates the sorts of considerations presented above (e.g., Simon 1979). Specifically, Simon famously argued that decision makers

typically satisfice rather than optimize.[88] A decision-maker normally chooses an alternative that meets or exceeds specified criteria, even when this alternative is not guaranteed to be unique or in any sense optimal. Simon argues that, instead of scanning all the possible alternatives, computing the probability of every outcome of each alternative, calculating the utility of each alternative, and thereupon selecting the optimal option with respect to expected utility, an organism typically chooses the first option that satisfies its "aspiration level". His concept of satisficing postulates, for instance, that an organism would choose the first object (a mate, perhaps) that satisfies its aspiration level instead of taking the time to survey all other possible alternatives. These considerations might thus be important when considering experimental results from, say, consumer or mate choice.

Following rule-based rationality might not just cost time. It might be quite effortful and painful as well. Recall that, according to rule-based rationality, people should conform to probabilistic norms such as Bayes' theorem—the idea that the probability of a hypothesis should be updated in light of new evidence by weighing both the base rate and the diagnostic evidence. Yet, as we saw in Chapter 1, people often neglect or underweight base rates. Let us now consider a case suggested by Elqayam (2012). Imagine that someone has lower than average intelligence. And imagine also that she is not motivated to engage in effortful processing, for she finds hard thinking to be particularly painful and not really productive. She might have goals that pull in different directions. On the one hand, she might want to give an accurate answer to a problem requiring the application of Bayes' theorem. However,

[88] It is worth noting that Simon also used the term for a specific heuristic: choosing the first alternative that satisfies an aspirational level.

on the other hand, she might also want to spend as little cognitive effort as possible. What is important to stress here is that, for a subject who finds hard thinking rather painful, it might be adaptively irrational to comply with rule-based rationality, as this would involve significant losses in time and huge and painful cognitive effort.

Notably, the abovementioned example suggests that AR theorists' concerns about the demandingness of rule-based rationality might apply to a broader set of cases than those offered by scholars appealing to the *ought-implies-can* principle, and to at least some tasks and examples from literature on MBR.

To conclude this section, let us briefly take stock of what we have achieved here: I have shown that there are ways of conceptualizing the normative significance of BR that do not appeal to versions of the *ought-implies-can* principle. In particular, from the perspective of AR, we have to take the bounds of cognition very seriously, as they characterize the epistemic contexts in which behaviour and cognition occur. Because of such bounds, following rule-based rationality might end up being maladaptive.

5. BR and optimism about human rationality

Therefore, it seems that, from the perspective of AR, some aspects of research on BR, and more precisely a person's time and cognitive limitations, do have normative relevance. In other words, given that these limitations are part of our epistemic contexts, following rule-based rationality might not be adaptive. Thus, it might seem

that the scenarios described above resemble the cases discussed in Chapter 3, where we encountered contexts in which, because of the particular configuration of the environment and the goals of the reasoners, following rule-based rationality turned out to be maladaptive, and violating rule-based rationality turned out to be adaptive.

It is worth reminding ourselves that appeals to the *ought-implies-can* principle had generally been used to question pessimistic views on human rationality: if following rule-based rationality is too demanding—it is often said—then its violations should not count as instances of irrationality. However, when we consider this alternative interpretation of the normative relevance of literature on BR, it becomes clear that we should be very careful when trying to draw optimistic conclusions.

To see this, consider that, if a reasoner is trying to make an accurate judgement in little time and energy, following a strategy that does not lead to accurate predictions but allows to save time and effort might be better than spending too much time and energy to get a more accurate prediction. However, if heuristics imply major losses in accuracy, then such behaviour cannot be seen as optimal or as a clear instance of adaptive behaviour and cognition.[89]

[89] Here we are suggesting a sort of ranking of different strategies according to their adaptive value. With regard to this point, it is important to highlight that the perspective of AR assumes that "rationality" is a comparative concept, and this marks a significant departure from traditional approaches towards rational behaviour and cognition. In particular, from the perspective of rule-based rationality, the assessment of behaviour is achieved quite straightforwardly by checking each norm and, if any are violated, irrational behaviour is claimed. The evaluation is binary because rule-based rationality is all-or-nothing in nature. As Morton pointed out, it is often claimed that 'intelligence is certainly a comparative concept, but some philosophers have denied that rationality is' (2012, 141).

To be sure, as we have seen in Chapter 3, AR theorists contend that fast-and-frugal heuristics might violate rule-based rationality and yet be successful, when measured in terms of their accuracy, speed and frugality. In Chapter 7 I will come back to these claims, examining whether these generalizations are warranted. Here, instead, I want to stress that what is often overlooked in the literature is that, in some cases, even small losses of accuracy might be quite important. Specifically, the importance of a goal seems to vary from context to context. For instance, the passage of time is certainly a pressing concern faced by an organism in a variety of dynamic environmental situations: organisms may have occasional speed-based encounters where the slower individual is placed at a serious disadvantage. Moreover, the faster an organism can make decisions and act on them to accrue resources or reproductive opportunities, the greater advantage it will have over slower competitors. But at other times, accuracy is definitely a more pressing concern. In particular, there are some decisions, such as whether to get married to someone, where making accurate decisions might be more important than saving time. In these latter sorts of contexts, even small losses in accuracy might result in maladaptive behaviour on the side of the cognizer. I take these considerations to suggest that, in some cases, saving time might not represent a significant benefit.

6. Conclusion

In this chapter, I have explored the connections between some aspects of BR and the normative framework of AR. In particular, I have shown that, whilst several scholars

in the "rationality debate" tend to conceptualise the normative relevance of both BR and the AR project by appealing to versions of the *ought-implies-can* principle, the framework of AR offers in fact an alternative way of doing so. I have also suggested that we take extra care in trying to draw optimistic conclusions about human rationality from considerations concerning human BR.

Chapter 6: Adaptive rationality meets research on individual differences

In this work, we have seen that biases can be instances of adaptive behaviour and cognition and that researchers should try to assess performance against the goals people entertain and against desired life outcomes. At this point, however, I also wish to discuss a reply that seems to be available to MBR theorists. These scholars can in fact appeal to other evidence to question the framework on rational behaviour and cognition articulated by AR theorists. In particular, in this chapter I consider the recent appeal to research on individual differences in reasoning and decision-making, which is mainly due to the work carried out by Stanovich. As it turns out, in several corners of research on judgement and decision-making it is claimed that such body of research has far-reaching implications for our understanding of human rationality and for the plausibility of the AR project. Here I reconstruct and discuss two different arguments based on this research which are directed at the AR project. First, heterogeneity in the use of heuristics seems to be at odds with the adaptationist background of the project. Second, the existence of correlations between cognitive ability and susceptibility to bias suggests that rule-based rationality is normatively adequate. I argue that, as matters stand, none of these arguments can be seen as fully compelling.

1. Introduction

In Chapter 3, we discussed and rebutted a number of moves that seemed to be open to MBR researchers to resist the conclusions by AR theorists. At this point, however, I wish to draw the reader's attention to a more general strategy to which advocates of

MBR research can appeal. Specifically, a group of researchers have argued that a crucial body of evidence has been unduly neglected by researchers in the "rationality debate" and, in particular, by AR theorists. The evidence I am referring to concerns findings on individual differences in reasoning and decision-making. Specifically, in a series of publications, Stanovich and his coworkers have argued that their reported findings ultimately undermine the AR project (e.g., Stanovich 2011b). The goal of this chapter is to assess whether Stanovich's arguments really undermine the AR project. The first argument I discuss is supposed to challenge the adaptationist underpinnings of AR. Precisely, Stanovich's reported findings on heterogeneity in the use of heuristics seem to be at odds with the idea that adaptationist pressures led to their use: one would expect their use to be far closer to universality if adaptationist pressures had led to them.[90] The second argument questions instead the normative claims made by AR theorists. The fact that people with higher cognitive ability follow rule-based rationality seems to suggest that rule-based rationality is normatively valid and there to stay, and that AR theorists should not try to replace rule-based rationality.[91]

I argue, however, that Stanovich's arguments fail to undermine the AR project. Not only commitments to adaptationism are not vital to the AR project, but I also argue that the actual heterogeneity of reasoning performance seems to be compatible

[90] My reconstruction of this argument follows Kelman's (e.g. 2013, 355). I take such reading to provide the strongest version of Stanovich's attack on the adaptationist background of AR (cf. Stanovich 2004, chap. 5).

[91] For instance, Stanovich writes that 'one aspect of this variability that researchers have examined is whether it is correlated at all with cognitive sophistication. […] We might take the direction of this association as a validation of the normative models' (2011b, 14).

with an adaptationist account, and perhaps even necessary for some plausible evolutionary stories. In addition, even the most plausible version of the second argument cannot be seen as fully compelling: it might be argued that people who score higher at tests of cognitive ability achieve better life outcomes because they do not reason heuristically. However, even if we grant that people with higher cognitive ability achieve better outcomes, the claim that they do so because they follow rule-based rationality remains at a hand-waving level and is not empirically well supported.

The chapter is structured as follows. In section 2, I discuss Stanovich's research on individual differences. In section 3, I reconstruct and assess the first argument based on his research. In section 4, I do the same for the second argument. In light of this discussion, I then conclude in section 5.

2. Stanovich's research on individual differences in judgement and decision-making

As we have seen in our investigation, research in the field of judgement and decision-making has described a variety of heuristics that reasoners seem to deploy (e.g., Gigerenzer and Goldstein 1996; Gilovich et al. 2002). Familiar examples are the *availability heuristic* (judge an event frequency by the ease with which instances of the event can be recalled; Kahneman and Tversky 1973) and the *recognition heuristic* (if you recognize only one item in a set, choose that one; Goldstein and Gigerenzer 2002). Yet, while several heuristics have been associated with human decision-making and formally modelled, little attention had been paid to the

existence of individual differences in their use until Stanovich and his co-workers (e.g., Stanovich 1999; Stanovich and West 2008) started to conduct a stream of individual differences studies involving reasoning and decision-making. A result of their research is that there is remarkable heterogeneity in the use of heuristics. Consider, once again, the *conjunction fallacy* (Tversky and Kahneman 1983). This phenomenon is usually interpreted as an indication of irrationality, because it violates the *conjunction rule* of probability theory. While most of the subjects in Stanovich's experiment displayed the conjunction effects, some did not (e.g., Stanovich 1999). Stanovich has pointed out that:

> What has largely been ignored is that although the average person might well display an overconfidence effect, underutilize base rates, violate axioms of probability theory, and so forth, on each of these tasks, some people give the standard normative response. (2011, 13)

There is systematic variability in all of these tasks: while people have been shown to have a strong propensity to use heuristics, not everyone does. In fact, a sizeable number of people do not deploy heuristics. Moreover, these people do not just randomly fail to use heuristics, but they systematically reason in a very different way from other humans. In the main, Stanovich has focused on the cognitive strategies invoked in the *heuristics-and-biases* tradition (e.g., Gilovich et al. 2002; Nisbett and Ross 1980), but large individual variability in strategy use has been reported also with regard to the heuristics generally modelled within the AR framework, like in the case of the *recognition heuristic* (cf. Richter and Spaeth 2006). The evidence available strongly supports a scenario where different types of reasoners, namely heuristic and non-heuristic users, coexist. Specifically, for several classes of

reasoning and decision-making tasks there are significant cross-task correlations: people that do not use heuristics in one context also do not do so in another (Stanovich and West, 1998, 2000; West, Toplak, and Stanovich 2008).

Stanovich's research, however, also shows that there are important correlations between the use of heuristics and cognitive abilities. It is useful to briefly introduce the concept of cognitive ability. When a diverse range of mental tests (e.g., understanding paragraphs, doing arithmetic, following instructions, estimating lengths, remembering words, identifying absurdities in pictures) is performed by a large group of people, the associations among the test scores form a pattern: no matter what type of mental work the tests involve, people who do well on one type of mental task tend to do well on all of the others. This phenomenon is known as general cognitive ability and it is usually shortened to just a lowercase italicized *g*.

Using standard measures of general cognitive ability, Stanovich and his colleagues examined effects that are among the most known in the literature, such as *base-rate fallacy, framing effects*, and *conjunction fallacies*, and the result of their research was that cognitive ability is associated with performance in those tasks. It seemed that people with higher cognitive ability were less susceptible to cognitive biases. But the accumulating findings have also resulted in some conflicting results. In particular, some evidence collected by Stanovich has more recently suggested that these associations between cognitive ability and performance in judgement and decision-making are not as strong as initially supposed. Some biases, such as *overconfidence* and *hindsight bias*, correlate with cognitive ability, but others, such

as *anchoring* and *sunk costs effects*, do not (e.g., Stanovich and West 2008). Moreover, while the existence of correlations between cognitive ability and use of fast-and-frugal heuristics proposed by AR theorists is undertested, some evidence suggests that in some contexts people who score higher at intelligence tests reason heuristically (Broder 2003).[92]

In light of this mixed evidence about the strength of the links between general cognitive ability and performance on judgement and decision-making tasks, Stanovich and his co-workers have also investigated other cognitive variables as predictors of performance. While cognitive ability might be a predictor of performance in judgement and decision-making, there is a better predictor, which is the so-called "cognitive reflection test" (CRT; Frederick 2005). Frederick has developed such test as a 3-item task shown to predict susceptibility to various cognitive biases. Stanovich and his co-workers have praised these measures as excellent predictors of performance in judgement and decision-making tasks (Toplak et al. 2011). Among the other cognitive constructs that have been explored, numeracy seems to hold promise for understanding and predicting behaviour (Reyna et al. 2009), where numeracy refers to the ability to understand and use numbers and has been shown to predict quite well susceptibility to a variety of biases and fallacies (Peters et al. 2006). While these other cognitive constructs might be better candidates for the prediction of cognitive biases, Stanovich has nonetheless highlighted that 'it

[92] Notably, however, Broeder (2003) used the Berlin Intelligence Structure (BIS) test, as it is common in German speaking countries. It is not entirely clear, however, how performance at this test relates to performance at other tests used to assess intelligence in Anglo-American research (Beauducel and Kersting 2002).

is never the case that subjects giving the non-normative response are higher in intelligence than those giving the normative response' (2011b, 15).

3. The argument from the heterogeneity in the use of heuristics

First, Stanovich's findings seem to be at odds with the AR project's adaptationist background and, more precisely, with attempts to understand the way people reason and make decisions by looking at human evolutionary history and appealing to the effect of natural selection. Specifically, Stanovich's reported heterogeneity in the use of heuristics seems problematic, as one would expect the use of heuristics to be far closer to universal if adaptationist pressures had led to their use. This argument is thus supposed to create troubles for the AR project's evolutionary underpinnings.

Notably, evolutionary psychologists often stress that the psychological mechanisms that evolved to solve Pleistocene adaptive problems now constitute 'an arrow of psychological mechanisms that is universal among *homo sapiens*' (Symons 1992, 139). Indeed, there seems to be a straightforward link between adaptationism and universality. The underlying reasoning seems to be that Pleistocene humans possessing a psychological mechanism that effectively solved an adaptive problem would have enjoyed a reproductive advantage over population members not possessing it, and that there was ample opportunity for selection to drive each beneficial psychological mechanism to fixation in early human populations.

Stanovich's argument can be summarized as follows. (1) There is heterogeneity in the use of heuristics. (2) If heuristics were the result of adaptationist pressures, then there would not be heterogeneity it the use of heuristics. (3) Heuristics are not the result of adaptationist pressures. The argument is thus based on two main premises; for the argument to go through, both of them have to be true. The truth of premise (1) can be granted relatively easily. Even AR theorists have accepted that individual differences in judgement and decision-making ought to receive more attention in future research, since 'in virtually every task we find individual differences in strategies' (Gigerenzer and Brighton 2009, 133).

Before I analyse premise 2, it is worth stressing that it is unclear whether the argument, even if sound, would turn out to be particularly damaging to AR theorists. To be sure, in Chapter 2 we did present some adaptationist considerations offered by AR theorists to account for people's violations of rule-based rationality in their reasoning and decision-making. But it is unclear whether a commitment to adaptationism is necessary for AR theorists.[93] With this clarification in place, now I want to argue that, whatever reason we may have to question an adaptationist perspective, this should not be based on Stanovich's research on individual differences in judgement and decision-making. Notably, this is only limited support for an adaptationist perspective. More specifically, whilst adaptationist considerations may offer some interesting suggestions or provide some support for some particular hypotheses, in Chapter 2 I also stressed the influence of factors

[93] One might want to emphasize that in fact AR theorists do not completely exclude the importance of factors beyond natural selection. In particular, it is interesting to note that, while Stanovich argues that AR theorists completely ignore the role of culture (2004, 132), at some times AR theorists do seem to appeal to cultural learning to explain people's behaviour (e.g. Hutchinson and Gigerenzer 2005; see also Arnau et al. 2014 for a discussion of the topic).

beyond natural selection, showing that they may result, for instance, in the evolution of non-adaptive traits and behaviour. The goal of this section is merely to show that there are moves available to adaptationists to accommodate Stanovich's findings, and I will briefly discuss some of the mechanisms that might result in actual heterogeneity. Moreover, I will show that a number of researchers have already started to link adaptationist models to actual heterogeneity in decision-making performance. Thus, I suggest that the argument reconstructed above does not work mainly because premise (2) is false.

3.1 Universally distributed heuristics and individual differences

Even if all humans exhibited no genetic differences, differences in the use of decision-making strategies could still occur as a result of different situational assessments. Consider the theory of evolutionary socialization, which seeks to establish a causal link between the perception of early childhood living conditions and later reproductive strategies (Belsky Steinberg and Draper 1991). According to the hypothesis, the degree of environmental stress experienced during childhood can be an indicator of adult reproductive conditions (e.g., the prevalence of monogamy and paternal investment). This hypothesis proposes that humans possess a conditional adaptation that uses childhood stress as a cue to channel maturation and psychosocial development so that they fit the demands of the predicted optimal reproductive strategy in adulthood. Individuals growing up in father absent homes during the first five to seven years of life develop the expectations that parental resources will not be reliably provided. Accordingly, insecure and unsupportive

family relationships would accelerate pubertal development and, in particular, females with such relationships would be enabled to initiate mating and reproduction earlier than females in secure and supportive family relationships. This effect would be advantageous in environments in which survival and thereby reproduction could be compromised. The opposite effect would occur in secure and supportive family relationships: pubertal development would decelerate, and females with such relationships would be able to delay reproduction and mating.

Moreover, conditional shifts can be due to adjustments to one's own physical phenotype. Tooby and Cosmides (1990) coined the term "reactive heritability" to describe evolved psychological mechanisms designed to take as input heritable qualities as a guide to strategic solutions. Any feature of the individual's world, including one's personal characteristics, that influences the successful attainment of those goal states may be assessed and evaluated by psychological mechanisms. Evolved mechanisms in this view are not early attuned to recurrent features of the external world, but can also be attuned to the evolution of the self. Suppose that all men have an evolved decision-rule of the form: pursue an aggressive strategy when aggression can be successfully implemented to achieve goals, but pursue a cooperative strategy when aggressions cannot be successfully implemented. Given this simplified rule, those who happen to be muscular in body can more successfully carry out an aggressive strategy than those who are skinny or endomorphic. If these individual differences in body build are at least partly heritable and provide input into the decision rule, they may produce stable individual differences in aggression

and cooperativeness that are adaptive and not directly heritable. They are rather based on a self-assessment of heritable information.

3.2 Heterogeneity in the distribution of heuristics and individual differences

Adaptive individual differences in the use of strategies might arise from genetic differences as well. In particular, balancing selection occurs when genetic variation is maintained by selection and might be a crucial mechanism maintaining genetic variation. The basic idea is that, if selection pressures vary over time or space, then selection may favour different levels of a personality trait in these different environments and consequently different strategies. For example, some environments might favour a risk taking personality, while others might favour a more cautious risk-averse personality. A recent natural experiment provides some support for the speculation that heterogeneity might be explained by appeal to balancing selection. Camperio Ciani et al. (2007) reported findings that indirectly support a role to balancing selection in sustaining genetic variance of personality. They studied average personality differences of Italian coast-dwellers compared to Italians living off the coast on three small island groups. They studied islanders from three different archipelagos isolated from each other. After matching populations for cultural, historical and linguistic background, and controlling for age, sex and education, they found that islanders from three distinct archipelagos isolated from each other share consistent, distinctive personality profiles. They differ from their respective mainlanders in being more conscientious and emotionally stable, and less extraverted and open. As expected, these differences were not very large; however, they were

significant and consistently observed in all the archipelago/mainland population pairs studied. This pattern makes cultural or developmental explanations for the population differences unlikely and suggests change on the genetic level. Even though individual fitness consequences of these traits were not measured directly, the apparent recent evolution of genetic differences between populations in these two traits suggests that the fitness payoffs of these two personality traits were historically distinct in these different environments.

Heterogeneity in the use of cognitive strategies might also be due to frequency-dependent selection, which is strictly speaking a particular case of balancing selection by environmental heterogeneity, concerning the composition of the social environment over space and time. Frequency dependent selection occurs when two or more strategies are maintained within a population at a particular frequency relative to each other, such that the fitness of each strategy decreases as it becomes increasingly common. For instance, Gangestad and Simpson (1990) tried to show that negative frequency-dependent selection could maintain heterogeneity in the context of mate choice. Their model is a very early and simplistic one, but it nonetheless provides a quite useful illustration. The central assumption is that women's mating strategies should centre on two qualities of parental mates: the parental investment a man could provide, and his genetic fitness. Yet, there may be trade-offs between selecting a man for his parenting abilities and selecting him for his genetic fitness: men who are highly attractive to women, for example, may be reluctant to commit to only one woman, resulting in a woman seeking a man for his genetic fitness having to settle for a short-term relationship without parental

investment. These two different goals are supposed to produce two alternative female mating strategies. Women seeking a high-investment mate are predicted to adopt a "restricted" sexual strategy marked by delayed intercourse and a prolonged courtship. Women seeking a man for the quality of their genes, on the other hand, have less reason to delay intercourse. Indeed, if the man is pursuing a short-term sexual strategy, any delay on her part may deter him from seeking sexual intercourse with her. Importantly, competition tends to be most intense among individuals pursuing the same mating strategy, and the two mating strategies of women—restricted and unrestricted—evolved and are maintained by frequency-dependent selection. As the number of unrestricted females in the population increases, the number of sexy sons also increases, and hence the success of the unrestricted strategy decreases. On the other hand, as the number of restricted females in the population increases, the competition for men who are willing to invest exclusively in them and their children increases, and the fitness of that strategy commensurably declines.

It is important to note, however, that appealing to frequency-dependent selection could be coming at a cost for AR theorists. Specifically, if AR theorists were willing to analyse human decision strategies in terms of optimal fit to some environmental factor, their attempt would then be (virtually by assumption) somewhat misguided—after all, people's strategies would then be shaped by the decision strategies other humans were using. However, this would only modify the details of AR theorists' account, leaving a general adaptationist project unscathed. As it turns out, adaptations, in the purest evolutionary sense, can be quite local. As soon as there

is an ecological niche, defined by selection pressures (which can well be the structure and characteristics of conspecifics, as in the case of frequency-dependent selection), some individuals can be fitter than others, and in that sense they are better adapted to this niche, even if it is spatially or temporally limited.

3.3 Concrete models for the evolution of heterogeneity

In this section I want to emphasize that, in fact, a number of scholars have already started to attempt to link evolutionary models to the heterogeneity in the use of strategies that is found in the lab (by appealing to frequency-dependent selection). The models I refer to seek to account for findings in behavioural economics, where people have been shown to exhibit individual differences in the ability to infer what other players will do, in their social emotions, and in guilt, anger, and reciprocity.

Consider the case of reciprocity and cooperative behaviour. People show consistent differences in their strategic approaches to cooperative economic games, with subjects exhibiting a range of strategies from completely trusting and trustworthy to tactical cooperation and free riding (e.g., Fischbacher et al. 2001; Fehr and Fischbacher 2003). For instance, while in public good games experiments one typically observes that people cooperate much more than predicted by standard economic theory assuming rationality and selfishness, observed cooperation is heterogenous (Goeree and Holt 2002). McNamara et al. (2009) have tried to account for these findings from an evolutionary perspective. Their suggested model predicts consistent variation between individuals in trustworthiness. Using evolutionary

simulations of an economic game, they suggested that individual differences in trustworthiness (i.e. cheating) could select for and maintain individual differences in trust, even if this brings a fitness cost. Notably, these results might well generalize to other traits, for example to costly social awareness and aggressiveness, and this study is particularly remarkable because it shows that the mere existence of socially relevant individual differences can foster the evolution of further individual differences.

Another interesting example concerns people's accuracy in the representation of what others are likely to do. An important result of research on behavioural game theory is that subjects do not seem to generally choose strategies as prescribed by mathematical game theory (Camerer 2003). Interestingly, an account that fits experimental findings quite nicely is "cognitive hierarchy" (CH; Camerer 2003) theory of strategic reasoning, which takes that there exists a hierarchy of player types, which corresponds to the different numbers of steps that players reason ahead in a game. CH modelling assumes that players have different levels of accuracy in their representations of what others are likely to do, which may vary from heuristic and naïve to highly sophisticated and accurate. Some people (level 0 players) just play at random. Other people (level 1 players) reason assuming that people play randomly in that way, and they then play the optimal strategy in light of this assumption, and other thinkers still (level 2 players) reason that some fraction of players are using a random strategy and that the remainder players are level 1 players, so they play the optimal strategy in light of this assumption, and so on. From an empirical and experimental perspective, most subjects seem to behave as type 1

and 2 players, and individuals of type 3 and above are quite rare (Costa-Gomes and Crawford 2006; Camerer 2003). Interestingly, Mohlin (2012) presented an evolutionary model of bounded strategic reasoning, stressing that an evolutionary process based on payoffs learned in these different games may lead to a heterogeneous population where most individuals belong to relatively low types. An important result of this evolutionary model is that it seems to offer support for CH and for the existence of bounded and heterogeneous theory of mind abilities.

4. Cognitive ability, individual differences, and heuristic reasoning

Stanovich's second argument against the AR project rests on his reported correlations between cognitive ability and people's susceptibility to bias and seems to be potentially more damaging to the prospects of the AR project, as it would call into question the very normative part of the AR challenge. The argument can be summarized this way: (1) people with higher cognitive ability follow rule-based rationality, (2) if people with higher cognitive ability follow rule-based rationality, then rule-based rationality is normatively adequate, and thus (3) rule-based rationality is normatively adequate. If the argument were correct, it would have important implications for the plausibility of the proposal I have been articulating so far. As it turns out, I have stressed that measuring behaviour against rule-based rationality is normatively problematic, whereas Stanovich stresses that rule-based rationality still offers the benchmarks of rational behaviour and cognition. Let us have a look at the proposal in more detail.

The plausibility of premise (1) depends on how we interpret it. On a broad understanding, the claim says that there is a strong correlation between performance on tests of cognitive ability such as IQ tests and CRT tests (but also numeracy tests), and the tendency to comply with rule-based rationality. In general, it seems true that these cognitive constructs supposed to measure people's ability to understand and solve different sorts of problems correlate quite well with people's susceptibility to biases. On a narrow understanding, however, the claim is that scores on standard intelligence tests correlate well with susceptibility to biases. This narrower claim is more problematic. As we have seen, it is true that negative correlations between IQ tests and susceptibility to bias have not been reported, but the evidence is still mixed: in some cases the connection is strong, in others it is tenuous (e.g., Stanovich and West 2008).

More controversial, however, is the plausibility of premise (2): how would the fact that people with higher cognitive ability comply with rule-based rationality provide support for the latter? This question raises a number of concerns. In making such an argument, Stanovich is accepting the view that intelligence is a stable trait and reasonably well measured by standard IQ tests, and that early-life manifestation of cognitive ability (i.e., IQ scores) follow people throughout their lives into better life outcomes. This view, however, is controversial. Few topics in psychology are as old and controversial as the study of human intelligence (see Mackintosh 2011 for a review) and several people—AR theorists included (e.g., Raab and Gigerenzer 2001)—contend that IQ tests do not measure anything at all. Here I want to make a non-trivial concession to Stanovich and leave well-known criticisms of cognitive

ability tests aside. Importantly, while this concession is not trivial, it is not unreasonable either. For instance, whilst critics of intelligence tests (e.g., Gardner 1983; Sternberg 1997) are eager to point out that there tests ignore important parts of mental life—many largely socio-emotional abilities, empathy and interpersonal skills—the predictive value of traditional cognitive ability tests has recently received important empirical support. It is quite difficult to state that those tests do not measure anything at all ('except performance on IQ tests', as people say), since test performance is strongly linked not just to education and income (Strenze 2007), but also to occupational success (Kuncel and Hezlett 2010), health (Gottfredson and Deary 2004), mortality (Calvin et al. 2010), as well as to various brain measurements (Jung and Haier 2007), and genetics (Haworth et al. 2010).

4.1 Cognitive ability and expert intuition

But why exactly would better cognitive ability have normative relevance here? I will briefly consider a first way to understand Stanovich's argument, which is in terms of an appeal to the intuitions of people with higher cognitive ability in order to justify norms of rule-based rationality. According to this reading, normative justification is based on intuitions and on the consensus of an *elite* of reasoners (see, e.g., Cokely and Feltz 2014 for a discussion of different appeals to experts' intuitions). We are thus looking at people with higher cognitive ability because their intuitions are more reliable, as they are more likely to track the truth, and this is the reason why their answers bear normative significance. In particular, Stanovich (1999; Stanovich and West, 2000) has appealed on some occasions to the understanding / acceptance

principle: the more one understands the normative principles involved in an inference task, the more likely one is to endorse these principles. This means that the more cognitively gifted reasoners are more likely to respond in congruence with the "appropriate" normative model—such as it is—for a particular problem set. Stanovich reverses the principle and maintains that we should accept as normative whatever is congruent with responses given by higher ability reasoners. Here the higher score on intelligence tests confers enhanced reliability to people's intuitions in virtue of their capacity to comprehend a particular case better.[94] On this view, tests of cognitive ability are important because they allow us to identify the most reliable intuitions. It thus seems clear that, while Stanovich has generally appealed to scores on IQ tests when substantiating his argument, the understanding / acceptance principle underlying claim (2) is consistent with a broad understanding of claim (1). In short, the fact that different operationalizations of cognitive abilities (not only those offered by IQ tests) correlate with susceptibility to biases matters because people who perform better in these different tests fare better at understanding and solving different problems. For instance, consider the case of numeracy tests, which is particularly interesting, since there have long been intelligence researchers who have speculated that mathematical ability is at the very core of general intelligence, and that the former sustains the ability to think clearly.[95] This application of the understanding / acceptance principle is not damaged by the fact that the "cognitive reflection test" is a better predictor of rational thinking than IQ tests.

---

[94] For instance, Stanovich writes that 'the direction that performance moves in response to increased understanding provides an empirical clue as to what is the normative model to be applied' (1999, 63).
[95] The idea that mathematical ability is at the heart of intelligence goes back to Vernon (1964). More recently, it has been shown that IQ and numeracy (mathematical ability) are both phenotypically and genetically correlated (Hart et al. 2009).

While this is a potentially interesting argument, I will not discuss it here any further. The appeal to the understanding / acceptance principle has been discussed and attacked in commentaries on Stanovich and West's (2000) BBS target article. Here I will just claim that there are important problems with it, even if we sidestep the traditional concerns with tests of cognitive ability.[96] There seems to be a problem of at least partial circularity in this justification: we know a normative system is adequate because the brighter participants comply with it, but some of the normative assumptions of those systems are incorporated into the intelligent tests used to determine who are brighter. In addition, the idea that intuitions can serve as a ground for normative justification has been criticized (e.g., Kahneman 1981; Baron 2000; Weinberg, Nichols and Stich 2001). It is often emphasized that what matters for the assessment of norms of reasoning and decision-making is their conduciveness to success in the real world. As we have shown in the introductory chapter, the idea that there is a crucial link between rationality and adaptive behaviour is widely shared in the literature on judgement and decision-making. Notably, Stanovich himself seems to believe that rule-based rationality ought to be justified pragmatically, by appealing to success in the real world (cf. Stanovich 2011a). For instance, Stanovich and West write that 'adaptive decision making is the quintessence of rationality', that 'to think rationally means taking the appropriate action given one's goals (instrumental rationality)' (2014, 81), and that 'a set of responses that leads to intransitivity hence leads to disastrous money pumps' (2004, 108). This results in this version of the argument ringing somewhat hollow: if what ultimately matters is success in the real

[96] Although discussing this evidence in detail falls beyond the scope of this work, it is worth stressing that a number of researchers have shown that experience and expertise do not always improve the quality of judgements and decision-making. For instance, Witteman and Tollenaar (2012) have shown that experienced mental health clinicians often do not outperform novices in diagnostic decision-making.

world, it is hard to see how an appeal to the better understanding of people with higher cognitive ability might help settle the debate, unless those people are also shown to be more practically successful.

4.2 Cognitive ability and success in the real world

I will now articulate a more promising version of the argument for the view that responses given by those with higher cognitive ability matter, which avoids charges of circularity and appeals instead to people's success in the real world. The basic idea is that we know that people with higher cognitive ability achieve better life outcomes, and the best explanation for this is that these people achieve these outcomes because they reason according to rule-based rationality and avoid biases. This would show that, *pace* AR theorists, following rule-based rationality is generally conducive to desired life outcomes, whereas following heuristics is not. While this argument has not been assessed so far in the literature, this formulation is still in line with Stanovich's pragmatic view of normative justification and with his idea that IQ tests are imperfect but nevertheless important measurements devices. Stanovich has explicitly stressed that cognitive ability tests are significant predictors of real life outcomes. According to him:

> Many evolutionary theorists have mistakenly downplayed cognitive constructs that are heritable (intelligence, personality dispositions, thinking styles) and that have demonstrated empirical relationships to behaviour that relate to utility maximization for the individual (job success, personal injury, success in relationships, substance abuse). (2004, 133)

Notably, the abovementioned quote might look odd, as expected utility theory actually makes no claim about the rationality of individual preferences. In fact, Becker and Murphy have developed some arguments that 'addictions, even strong ones, are usually rational in the sense of involving forward looking maximization with stable preferences' (1988, 675). There is, I think, a clear way to make sense of Stanovich's claim: as a matter of fact, most people do care about job success, health, and other life outcomes. This does not mean that we can easily come up with a clear ranking of such life outcomes. When deciding to buy, say, insurance, the problems related to such a ranking might come out quite clearly. For instance, health is good, but fun is good too. This is an important caveat, but I still think that we can grant that the goods Stanovich refers to are widely regarded as good life outcomes, and that it is interesting to ask whether cognitive ability is a good predictor of those, and what, in case it is, mediates its effect on these outcomes.[97]

My point here is that, even if we accept that people with higher intelligence achieve better life outcomes, the rub comes when one seeks to explain this by appealing to the fact that these people reason in compliance with rule-based rationality. In part, this is because these correlations are more tenuous than Stanovich might want: it is not true that IQ is always such a strong predictor of susceptibility to bias. Other operazionalizations of cognitive abilities are stronger predictors (such as the cognitive reflection test), but since the well-established correlations between cognitive abilities and real life outcomes are based on IQ scores, the evidence we

[97] Here I am not claiming that these are normatively and objectively "desirable" outcomes. Otherwise, this would actually look like a possible *reductio* of the perspective of "instrumental rationality" that I have accepted in this work. Here I am just claiming that, even if what matters is that these outcomes are subjectively desired, it seems that behaviour could still be usefully assessed against such goals, as people tend to treat goods such as health as important.

need to take into account is about the correlations between IQ scores and susceptibility to bias. Yet, there are even more serious difficulties here, as Stanovich's research might just appeal to correlations and does not bring out the relevant causal connections. Thus, we lack the needed details to fully secure the causal story, and his hypothesis rests on little more than sketchy speculation.

To show that these concerns are not at all untethered, I will focus on the discussion of the domain of health. The reasons for analysing this domain is that it will make the discussion of these key problems more tractable and that it has been the focus of a great deal of interest from AR theorists as well (e.g., Gigerenzer and Gray 2011). This discussion can serve as something of a cautionary tale about the care that must be taken in developing causal explanations of associations between cognitive ability and success.

The field of study of intelligence and health outcomes is called "cognitive epidemiology" (e.g., Deary, Weiss and Batty 2010). Cognitive epidemiology grew out of, essentially, Deary's conception of cognitive ability, which is strongly rooted in the tradition of thinking of intelligence as a stable, genetically influenced trait that is measured well by IQ tests. More precisely, a lot of work has been dedicated to providing evidence that early-life manifestation of IQ follow people throughout their lives into better health outcomes. Intelligence was inversely related to the risk of alcohol disorders, depression, generalized anxiety disorder, and posttraumatic stress disorder (Gale et al. 2008). But intelligence is also strongly correlated with physical health. For instance, the hazard of being involved in a fight or a brawl is over eight

times greater for the lowest versus highest IQ group, and an elevated risk of lung cancer has been reported in adult men and women who had lower intelligence test scores in childhood (Hart et al. 2003).

One might want to posit the following story as an explanation of the correlations between IQ and health outcomes: more intelligent people achieve better health outcomes because they avoid biases. After all, health self-care is a complex set of tasks that require knowledge, decision-making, planning and engagement. It is thus tempting to argue that one's level of cognitive ability will be related to health outcomes and ultimate survival because of the better understanding and decision-making (Gottfredson and Deary 2004). More intelligent people have better health and numeracy literacy, and are thus less vulnerable to biases. It is quite clear that poor health literacy and numeracy can come at high cost. In the words of Gottrfredson, 'misreading a map or a train schedule may be nuisance, but misreading a prescription label might be an hazard' (2004, 180). Also, the effectiveness of cancer treatments is expressed as survival rates (e.g., the percentage of treated patients who survive for five years), the benefits of lifestyle changes as reductions in cardiovascular risk, and the side effects of medications as probabilities of death, discomfort, and disability. Thus, the story goes, people with higher intelligence manage health care more effectively, and ultimately achieve better health outcomes because they tend to engage in rational thinking and avoid biases. People with lower intelligence might be victim of *framing effects*, *overconfidence*, and other sorts of biases, and this explains the achievement of worse health outcomes. Now, I do not want to argue that the idea that differences in reasoning and decision-making explain

different health outcomes is somewhat misguided or could never be made to work, but rather that the posited story lacks the needed support.

To assess the plausibility of the hypothesis we need to consider both the empirical evidence available and the rival hypotheses. Notably, there is slim evidence for the claim that people with higher cognitive ability achieve better outcomes *because* they are less susceptible to biases. Not many studies have investigated the mediators of the impact of intelligence on health behaviours. For instance, Di Matteo (2004) found that higher income and education were associated with better compliance, but the review lacked information about intelligence. Interestingly, Deary et al. (2009) have investigated whether higher intelligence predicted long-term compliance with medication for up to two years in individuals who knew themselves to be relatively at high risk of cardiovascular disease, finding some important effects. Moreover, supporting evidence from self-reports shows that people with higher childhood intelligence tend to exercise more and have diets that accord better with health information, are less likely to smoke or to be obese or overweight, and have fewer hangovers from drinking alcohol (Batty, Deary and Macintyre 2007). Yet, while these studies might show the impact of intelligence on health behaviour, they are not informative about whether this is due to compliance with rule-based rationality. What is needed here is evidence that what mediates the impact of intelligence on desired outcomes is the tendency to comply with rule-based rationality.

Moreover, there are alternative frameworks and explanations for the predictive value of intelligence and IQ tests, which do not appeal to different strategies in

reasoning and decision-making. For instance, intelligence is associated with more education, and thereafter with more professional occupations that might place the person in healthier environments. It might be that more intelligent people get better jobs, and that disparities in material resources, work environment, and access to medical health care result in different health outcomes. According to this view, it is still the case that intelligence is responsible for different health outcomes, but material resources are posited as mediators of this influence. Alternatively, brighter people might be "hardier" and have stronger immune systems genetically to begin with, so they can take more "lifestyle health abuse" along the way without ill effects (Deary 2012). While controversial, this is a fascinating and interesting hypothesis: health and cognitive ability are seen as two indicators of an individual's system integrity. The theory of general body integrity posits that the inverse association between premorbid cognitive ability and health outcomes can be explained by underlying physiological make up, and childhood mental ability is an indicator of system integrity.[98]

The issues touched upon in the previous paragraphs are obviously intricate and there is more to be said than space allows for here. What emerged, however, is that there are different hypotheses about the role of intelligence for the achievement of desired life outcomes. There is no clear theory yet as to what really explains the impact of intelligence on health outcomes. While there is no reason why different hypotheses cannot co-exist, the main problem stems from the difficulty of disentangling them and establishing their relative importance. Quite clearly, Deary

[98] Interestingly, this view seems to support Juvenal's centuries-old dictum to pray for a healthy mind in a healthy body (*mens sana in corpore sano*).

claims that 'although intelligence plays a part in health behaviours and health outcomes that contribute to specific causes of death, a clear chain of causation from intelligence to health outcomes and then to death has not emerged' (2008, 456).

Thus, as matters stand, we should refrain from inferring that compliance with rule-based rationality is what explains the achievement of successful outcomes by people with higher cognitive ability. But this is exactly the causal story Stanovich has to provide to get his argument off the ground. The verdict of this section, however, has been merely that the argument cannot be seen as compelling until the hypothesis becomes further substantiated. This is not, however, a structural limitation of the argumentative strategy, and this section should be seen as a call for more, not less work, in the attempt of disentangling different causal hypotheses and explaining what mediates the impact of intelligence on desired life outcomes. If it turned out that judgement and decision-making that conforms to rule-based rationality makes a major contribution to the achievement of desired life outcomes, then there would be some interesting evidence at odds with tenets of the AR project.

5. Conclusion

In this chapter, I have tried to show that Stanovich's appeal to his research on individual differences in reasoning and decision-making to undermine the AR project is not particularly compelling. First, I argued not only that commitments to adaptationism are not vital to the AR project, but also that the actual heterogeneity of reasoning performance is compatible with an adaptationist account, and perhaps even

necessary for some plausible evolutionary stories. Second, the existence of correlations between measures of cognitive abilities and susceptibility to biases does not provide much support to rule-based rationality. In particular, the hypothesis that more intelligent people achieve better outcomes because they comply with rule-based rationality lacks the needed support.

Chapter 7: Adaptive rationality, biases, and the *divide et impera* strategy

As we have seen, AR theorists have shown that biases reported in the MBR literature can be instances of adaptive behaviour and cognition, and that researchers should try to assess performance against the goals people entertain. Moreover, in the previous chapter I also sought to show that AR theorists' challenge is not, as yet, scathed by recent research on individual differences in reasoning and decision-making. But does this mean that we should accept AR theorists' optimistic claims about human rationality? In this chapter, I show that there are reasons to be sceptical about such claims. In the main, here I focus on a new strategy to resist AR theorists' revolutionary rhetoric by taking issue with their claim that the biases reported in the MBR literature should be conceived of as violations of rule-based rationality. As I try to show here, several important families of biases are not just violations of rule-based rationality, and do not seem to be clearly vulnerable to the challenge mounted by AR theorists. In fact, it seems that many biases reported in the literature have been assessed against prudential and epistemic goals, and that these findings do not sit well with AR theorists' claims about human rationality. I argue that AR theorists have not made clear how, in light of such evidence, they could hold optimistic views about human rationality.

1. Introduction

In Chapters 3, 4 and 6 I discussed the plausibility of the AR project from a normative standpoint. In this chapter, I will look at the AR project from a different perspective. Specifically, the AR project has often been presented as attempting to show 'how

people are able to achieve intelligence in the real world' (Todd and Gigerenzer 2012, 3), and provocative book titles such *Adaptive thinking: Rationality in the Real World* (Gigerenzer 1999) or *Ecological Rationality: Intelligence in the Real World* (Todd and Gigerenzer 2012) illustrate quite clearly the general aims of their project.[99] But are AR theorists' optimistic claims about human rationality warranted? Here I raise some problems for the claim that the acceptance of the AR's normative challenge licenses optimistic verdicts on human rationality.

This chapter details my reply to AR theorists. First, in section 2, I present some initial considerations that suggest that we should avoid overly optimistic conclusions about human irrationality. In particular, I suggest that, at least in several contexts, violations of rule-based rationality correspond also to failures from the perspective of goal-based rationality. Later, in sections 3-7, I substantiate my *divide-et-impera* move by discussing families of biases that do not seem to be vulnerable to the challenge to rule-based rationality mounted by AR theorists and, at least *prima facie*, look instead like cases of unsuccessful behaviour assessed against epistemic and prudential goals. In section 8, I rebut three objections, and in section 9 I discuss the relationship between my conclusions and evolutionary considerations. Section 10 provides a summary and a conclusion.

[99] Moreover, Gigerenzer argues that we need not worry about human rationality (1997, 280), that 'the rationality of the adaptive toolbox is not logical, but ecological' (Todd and Gigerenzer 2012, viii), and that 'often what looks like a reasoning error from a purely logical perspective turns out to be a highly intelligent judgment in the real world' (Gigerenzer 2008, 107).

2. Rule-based and goal-based rationality do not always diverge

First, I would like to stress that it is not clear to what extent assessments based on rule-based rationality might diverge from those relying on goal-based rationality. As we have seen in Chapter 3, in a number of contexts violations of rule-based rationality might be instances of successful behavior from the perspective of goal-based rationality. In light of this evidence, AR theorists tend to glorify heuristic reasoning, and some of their studies showing that heuristics violating rule-based rationality lead to successful behavior have gained a sort of mythical status. Yet it is still unclear to what extent this effect might be generalized, and there are reasons to think that, at least in several contexts, violations of rule-based rationality correspond also to failures from the perspective of goal-based rationality.

To see this, consider the evidence on fast-and-frugal heuristics presented in Chapter 3. It is true that fast-and-frugal heuristics that neither look up nor integrate pieces of information might be as accurate as (and even outperform) compensatory strategies that do not violate transitivity. However, while one is probably convinced when reading AR theorists' material that strategies like the *recognition heuristic* are efficacious in some situations, it is also natural to worry about how this relates to many other contexts. The reader will probably recall from Chapter 3 that the yes/no recognition response is used by the *recognition heuristic* as a frequency estimation cue: if one of the two alternatives is recognized and the other is not, it should be inferred that the recognized alternative has a higher value. In a famous experiment, Americans and Germans had to find out which was the more populous city between San Diego and San Antonio (Gigerenzer and Goldstein 1996). Many Germans

recognized the former but had never heard of the latter, and all of them chose the former over the latter, potentially relying on the *recognition heuristic*. All the Germans gave the right answer and performed better than the more knowledgeable Americans (who recognized both the cities and, therefore, were not able to use the *recognition heuristic*).[100] What I would like to stress, however, is that if AR theorists had run experiments asking people, for instance, to choose which of two large (but not extremely large) Asian cities is more populous, or asked subjects to compare the size of nearby tiny towns and unknown state capitals, these scholars would have probably demonstrated a sort of *recognition bias*, which would look like an instance of maladaptive behavior discovered through goal-based rationality. Moreover, if subjects were to assess the relative population size of different animals or the relative safety of two airlines based on recognition, the correlation between recognition and the criterion value would be in fact weak (cf. Richter and Spath 2006). For a clearer illustration of this point, consider the paper by Borges et al. (1999), which aimed to compare different stock-picking strategies. The fast-and-frugal strategy was simply the strategy of picking the most familiar companies (an application of the *recognition heuristic*). The experimenters constructed portfolios of companies recognized by "laypersons" selected at random from passers-by in Chicago and Munich and compared the performance of these portfolios against the comparison strategies. Given that for the six-month period of study the *recognition heuristic* outperformed the other strategies, the authors have suggested that ordinary people (who are disposed to naturally use the *recognition heuristic*) can perhaps do better on the stock market than mutual fund managers and market indices. All this resulted in an arousal

[100] As I stressed in Chapter 3, the answer was correct at the time of the study, but it would not be correct anymore.

of enthusiasm for the *recognition heuristic*, because this simple algorithm seems to perform well in real-life. However, there is some evidence that the success of the *recognition heuristic* was due to some luck, because the study by Borges et al. (1999) was carried out during a historically strong rising market. Boyd (2001) tested the *recognition heuristic* on a bear falling market, finding that the simple rule gave below average returns this time.

This is just to illustrate that, while the heuristics described by AR theorists might be successful in a number of cases and contexts, in other cases they seem to lead to unsuccessful behavior.[101] AR theorists seem to overlook these downsides, and thus they seem to have oversold their case in favor of human rationality. In other words, there are reasons to think that, for a number of biases that are violations of rule-based rationality, behavior that exhibits such biases will also be a failure of goal-based rationality.

3. Taking the descriptive issue seriously: the *divide et impera* strategy

But the reader might wonder, at this point, whether there are not other and perhaps more convincing strategies we could adopt in trying to resist AR theorists' most ambitious and optimistic claims about human rationality. In this section, I will detail a stronger reply to AR theorists. Specifically, I try to show that several areas of MBR do not seem to rely merely on the normative perspective of rule-based rationality. In fact, it turns out that different criteria of accuracy have been used in the scientific

[101] Note, however, that here I am not claiming that it is clear how to get good outcomes by applying rule-based rationality to the stock market. I am simply stating that the merits of fast-and-frugal heuristics in this context have been probably overstated.

study of reasoning and decision-making. In light of these considerations, I suggest that several families of biases do not seem to be vulnerable to the challenge to rule-based rationality mounted by AR theorists. In fact, I also show that several important families of biases have been assessed against epistemic and prudential goals and argue that AR theorists have not explained clearly why these biases should not be considered to be instances of adaptively irrational behaviour or how these findings could be accommodated within an optimistic framework on human rationality. Since my strategy consists in emphasizing that the category of bias refers to a range of heterogeneous phenomena and violations of different standards of accuracy, I will refer to as the *divide-et-impera* strategy. In offering this strategy, I draw attention to a set of overlooked descriptive issues.[102]

In what follows, I articulate and substantiate this reply to the AR project. It is worth highlighting that AR theorists are eager to claim that MBR has relied on norms of logic, probability theory, and rational decision theory or, to use our own term, rule-based rationality. For instance, AR theorists Wilke and Mata claim that in MBR:

> Participants were presented with a reasoning problem to which corresponded a normative answer from probability theory or statistics. Next, participants' responses were compared with the solution entailed by these norms, and the systematic deviations (biases) found between the responses and the normative solutions were listed. (2012, 531)

As the reader will recall, AR theorists are certainly not alone in providing such a characterization—this picture is quite widely accepted. For instance, psychologists

[102] Notably, other authors have suggested that biases seem to be heterogeneous (e.g., Arkes 1991; Stanovich 2011). But while they argue for this claim by pointing to the different cognitive processes involved in different types of biased reasoning, I focus on the evaluative standards against which such biases have been assessed.

Nisbett and Ross write that they 'follow conventional practice by using the term "normative" to describe the use of a rule when there is a consensus among formal scientists that the rule is appropriate for the particular problem' (1980, 13). Alternatively, consider Baron, who writes that 'the major standards come from probability theory, utility theory, and statistics. These are mathematical theories or "models" that allow us to evaluate a judgment. They are called normative because they are norms' (2004, 19). Moreover, Tversky and Kahneman (1986) describe their project as relying on the normative rules of 'the modern theory of decision making under risk' (S252). They consider *transitivity* of preferences, *dominance*, *invariance*, and *cancellation*. Overall, it seems quite clear that, if we look at the literature on judgment and decision-making, researchers in different groups agree that the biases reported in the psychological literature are violations of rule-based rationality and that have been identified by relying on such a normative perspective.

Whilst the descriptive claim that MBR relies on rule-based rationality has generally been accepted at face value, it is also quite inaccurate. It should be stated clearly, here, that I do not mean to deny that the characterization of biases as violations of rule-based rationality nicely fits some parts of research on judgment and decision-making, and we have discussed quite in detail some of these cases in this work. For instance, the influential Wason selection task presented in Chapter 3 was invented to explore the degree to which human thinking matches the laws of deductive logic and propositional calculus (Wason, 1966). And consider violations of the *conjunction rule* of probability theory or of the *axiom of descriptive invariance*, which have attracted a great deal of attention in the *heuristics-and-biases* tradition.

Wason selection tasks, *conjunction fallacies* and other overdiscussed studies have been taken to be representative of the phenomena investigated in MBR. But this is a misleading trend, or so I contend.

It turns out that the characterization of MBR as entirely relying on rule-based rationality is exceedingly narrow, and that there are many important families of biases documented in both cognitive and social psychology that do not seem to fit well with such characterization.

Before I demonstrate this, it is worth considering once again the words of AR theorists. According to Gigerenzer and Sturm, 'while ecological rationality is broadly defined in terms of success, and thus involves looking for means suited to certain goals, we do not maintain that reasoning is only about satisfying desires, without caring what is actually true or correct' (2012, 255). Now, interestingly enough, in many areas of research on judgement and decision making inaccurate behaviour has been assessed specifically against epistemic and prudential goals.

More precisely, in many areas of research on biases, researchers have not assessed behaviour by relying on rule-based rationality alone: in the case of several important families of biases, violations of rule-based rationality are neither necessary nor sufficient conditions for their occurrence. In fact, what seems to constitute a number of biases is that reasoning and decision-making violate other criteria of accuracy. Here I want to suggest that AR theorists are targeting a limited group of biases.

Notably, a small number of researchers seem to have noticed this point. For instance, Kruglanski and Ajzen argue that:

> Contemporary research on bias and error in human judgment is decidedly empirical in character. It lacks a clearly articulated theory and even the central concepts of "error" and "bias" are not explicitly defined. Nor is it easy to find a clear characterization of the objective, or unbiased inference process from which lay judgments are presumed to deviate (Kruglanski and Ajzen, 1983, 2)

In the remainder of this chapter, I wish to show that in several areas of MBR scholars do not seem to have heavily relied on rule-based rationality. I will focus, in particular, on a number of biases that do not seem to have been identified using rule-based rationality or to be constituted by violations of rule-based rationality. Some of the families of biases that I will discuss go under the headings of *overplacement*, *overestimation*, and *mental contamination*. I want to emphasize that the phenomena in which I am interested are not minor ones, but rather main effects explained in terms of biases and associated with irrational behaviour in research on judgment and decision-making. Moreover, at least *prima facie*, the biases I consider look like instances of unsuccessful behaviour measured against standard epistemic goals and people's individual desires and preferences. Besides these findings, I will also point to recent research on happiness psychology and to research on the detection of lies, where behaviour and cognition have been seemingly assessed against people's goals. In light of my discussion, I suggest that AR theorists have not as yet explained why these instances of behaviour and cognition should not count as plausible cases of irrational behaviour from the perspective of AR and why these findings should not be taken to be worrying. After this, I will also discuss some possible replies coming from AR theorists.

But before I start with my presentation of the *divide-et-impera* strategy, I wish to emphasize that there are several other areas of research that bear on the plausibility of the AR theorists' position on human rationality and that have been widely overlooked.

In particular, there has recently been a great deal of interest in the science of false memories (Brainerd and Reyna 2005). With false memories, one remembers events as having happened at some moment in one's life, although in fact the particular events did not happen then, or ever. False memories can be harmless—as when one thinks she has served a bottle of Sauvignon and it was actually Gewürztraminer. In fact, some scholars have also argued that false memories can offer important benefits for an agent (e.g., Fernandez 2014; Conway and Loveday 2015). But there are also many other circumstances in which such errors seem to be far from beneficial. Consider, for instance, a doctor prescribing treatments based on false memories of the symptoms of patients (Reyna and Lloyd 1997). Yet, it is in the legal realm, where one meets perhaps the most frequent cases of harmful false memories. Not surprisingly, studies of known cases of false conviction have suggested that a number of prototypical forms of legal false memory, like in the case of false identifications of innocent suspects (e.g., Wells et al. 1998) and false recollections during interrogations (e.g., Kassin 2005), are the leading cause of false convictions (Brainerd and Reyna 2005). These seem to be cases of factually "inaccurate" judgments in which memory judgments are evaluated in terms of correspondence with past events. More should be said about the costs and benefits of these kinds of

inaccurate judgements. However, what I want to highlight here is just that there are many more findings than the ones I can cover in this work which seem to have bearing on the assessment of people's rationality and on the plausibility of AR theorists' conclusions.[103]

4. Mental contamination

I will start to substantiate the *divide et impera* strategy by introducing research on so-called *mental contamination*. Wilson and Brekke distinguished between two different types of biases. Specifically, according to them:

> In the last 20 years, cognitive and social psychologists have documented numerous errors in reasoning, bias in judgment, and flawed heuristics. […] We suggest that the numerous instances of biases in human judgment are of two general types: the failure of rule-knowledge or application and those that result from mental contamination (cases whereby a judgment, emotion or behaviour is moved by unconscious or uncontrollable processes). Wilson and Brekke (1994, 118)

The distinction offered by Wilson and Brekke (1994) is probably not exhaustive, as we will see in the following sections, but it is nonetheless useful for the present discussion. Cases of *mental contamination* have generally been taken to be biases and instances of irrational behaviour in MBR, but it does not seem that researchers interested in such phenomena have heavily relied on rule-based rationality in the identification of such flawed behaviour. In particular, it does not seem to be the case that what constitutes these biases is that they are violations of rule-based rationality.

[103] For other lines of research in which people's judgements have been assessed against empirical accuracy and seem to be quite often empirically inaccurate, see, e.g., Perilloux (2014; Perrilloux et al. 2012).

What research on mental contamination has shown in the past three decades is rather that people are often influenced by such unconscious and unwanted factors (unwanted by the subject, not by others or by society) in a broad range of contexts.

4.1 Mental contamination and implicit biases

To appreciate one particular manifestation of mental contamination, we should consider the extensive literature on the *implicit bias*. In the words of Brownstein:

> *Implicit bias* is a term of art referring to relatively unconscious and relatively automatic features of prejudiced judgment and social behavior. While psychologists in the field of 'implicit social cognition' study 'implicit attitudes' toward consumer products, self-esteem, food, alcohol, political values, and more, the most striking and well-known research has focused on implicit attitudes toward members of socially stigmatized groups, such as African-Americans, women, and the LGBTQ community. (2015)

One of the remarkable features of *implicit bias* is that people might not be aware of being influenced. Interestingly, the *implicit-association-test* has been used to show that a great many people who profess to be racially impartial and explicitly disavow any form of racial prejudice display signs of racial bias in controlled experimental settings. Such biases may affect not only our beliefs, but also the way we perceive the world. A study by Payne showed that participants were able to more quickly identify guns (as opposed to non-gun tools) when they were primed with a black face, as compared to when they were primed with a white face. In this experimental set up, the 'presence of Black faces facilitated the identification of guns relative to the presence of White faces' (2001, 185). In these cases, it might be that unconscious

and unwanted processes influence people's cognition. If this situation occurs, the resulting bias looks like an instance of mental contamination and not, strictly speaking, a clear example of a violation of rule-based rationality.[104]

4.2 Mental contamination and anchoring

Moreover, consider now the following seminal study on the so-called *anchoring effect*. A wheel of fortune is spun and stops at the number 65. You are then asked if the percentage of African countries in the United Nations is above or below that number. Could this exercise actually influence your estimate of the percentage? Although it may seem unlikely, the evidence is that such anchors have an effect: in fact, groups who received larger numbers determined by a wheel of fortune gave higher estimates than groups who received lower numbers, demonstrating that irrelevant anchors influenced such estimates (Tversky and Kahneman 1974). Yet these numbers are in no plausible way related to the actual number of African countries. The idea, then, is that people's behaviour is biased by unconscious factors. Moreover, it seems that these factors are also unwanted, as they interfere with the subjects' goal and attempt to make accurate predictions. This is important: while talk about percentages in this example might suggest that MBR scholars focusing on anchoring are just relying on rule-based rationality, I want to emphasize that researchers trying to identify these biases have been seemingly relying on goal-based rationality. Specifically, *anchoring* refers to a rather general phenomenon in which people's attempts to give accurate estimates (e.g., about the height of a tree, the cost

[104] It is worth noting that the extent to which implicit biases are unconscious is controversial (e.g., Hahn et al. 2014). For a critical discussion of literature on the effects of unconscious factors on decision-making, see Newell and Shanks (2014).

of a house or the length of a river) are supposedly compromised by automatic and unconscious processes. The claim that scholars in MBR have just relied on rule-based rationality in their assessment of behaviour and cognition can be challenged, by pointing to these situations in which people simply make inaccurate predictions.

What seems to emerge by considering these important families of biases is that the characterization of MBR in terms of an assessment of behaviour and cognition against rule-based rationality is just too narrow and inaccurate. By considering cases of *mental contamination*, I have suggested that several biases have not been identified relying on rule-based rationality, and that they are not necessarily constituted by violations of rule-based rationality. This suggests, I think, that these families of biases are not clear targets of the AR theorists' challenge. In fact, at least some of them seem to be plausible cases of unsuccessful and irrational behaviour from the perspective of AR, since there is a clear appeal to the agent's goals in the individuation of each bias. Specifically, for such biases to occur, automatic and unconscious factors need to interfere with the agent's fulfilment of her desires and goals. And what is particularly interesting, here, is that a pillar of AR is the very the idea that we should take into account the agent's goals and preferences when assessing her behaviour.

5. Flawed self-assessments

I will now move on to a different line of research on biases. More precisely, I will now consider the literature on flawed self-assessment. This strand of research does

not seem to rely on rule-based rationality either. In these cases, subjects' unsuccessful behaviour is measured against empirical accuracy. In a much-cited paper, Taylor and Brown (1988) show how people's beliefs seem to lack an objective grasp of reality: in the cases reviewed by these authors, people's judgments and predictions about themselves and the world are systematically flawed. When it comes to this kind of mental error, we are dealing with inaccurate beliefs about the world. This literature on flawed self-assessments is huge. Sometimes, this extensive literature is referred to using the general term of *overconfidence*, but we should be more precise about the nature of this kind of bias. Here, I will use the terminology introduced by Moore and Healy (2008, 52) and distinguish between *overestimation (*an overestimation of one's actual ability, performance, level of control, or chance of success) *overplacement* (where people believe themselves to be better than others, such as when a majority of people rate themselves better than the median), and *overprecision* (excessive certainty regarding the accuracy of one's beliefs). It is interesting to note that, while AR theorists often discuss research on *overprecision*, they have not paid much attention to cases of *overestimation* and *overplacement*. This is important: *overprecision* concerns people's confidence in the accuracy of their judgments, and not the accuracy of their judgments. In other words, investigating whether my belief is true or false is different from investigating how confident I am in that belief.

5.1 Overestimation

One class of flawed self-assessments comes under the heading of *overestimation*. People seem to overestimate their skills, beliefs and performance. Consider the *planning fallacy* (Kahneman and Tversky 1979), which is the tendency to hold a confident belief that one's project will proceed as planned, even while knowing that the vast majority of similar projects have run late. As Buehler, Griffin and Ross point out:

> Anecdotal evidence of the *planning fallacy* abounds. The history of grand construction projects is rife with optimistic and even unrealistic predictions, yet current planners seem to be unaffected by this bleak history. (2002, 252)

A great deal of evidence has been offered to demonstrate that people consistently overestimate how easily they can complete a task (as measured by time or money) (e.g., Buehler, Griffin and Ross 2002). For example, the amount of time students take to finish their senior thesis is three weeks longer than their most realistic estimate of how long it will take—and a week longer than what they describe as their worst case scenario.

Another interesting case of inaccurate prediction is offered by the so-called *illusion of control*, which occurs when people overestimate their control over events. In a series of experiments, Langer and colleagues (1975) found that people often act as if they had control in situations that are actually dominated by chance. But similar

effects have been replicated in several contexts (e.g. Thompson, Armstrong and Thomson 1998).

The abovementioned biases seem to be instances of factually erroneous beliefs. In these cases, behaviour has been assessed against the goal of empirical accuracy, under the assumption that people are trying to make empirically accurate judgements. This suggests quite clearly, I think, that biases such as the *planning fallacy* are not sensitive to the challenge on rule-based rationality mounted by AR theorists. In fact, since AR theorists generally claim that we should measure behaviour against epistemic goals and prudential goals (they themselves have actually quite often assessed people's behaviour against empirical accuracy), and the behaviour discussed above seems to involve empirically inaccurate beliefs, it is not clear how AR theorists could, in light of such evidence, hold their optimistic claims about human rationality.

5.2 Overplacement

When people have unrealistic views about their position with regard to others, they are also subject to *overplacement*. Consider for instance the *better-than-the-average* effect. The classic example of this tendency is a 1981 survey of automobile drivers in Sweden, in which almost 90% of the people described themselves as above-average drivers. These effects have been shown to be widespread: for instance, motorcyclists believe that they are less likely to cause an accident than the typical biker (Rutter, Quine and Albery 1998). Moreover, business leaders tend to believe that their

company is more likely to succeed than is the average firm in the industry (Cooper, Woo and Dunkelberg 1988). People tend to have similarly unrealistic views about their position with regard to others also when they look at the future. People are in fact prone to the *optimism bias*. They seemingly perceive of their future as more positive than that of the average person. Most college students, for instance, tend to believe that they will have a longer-than-average lifespan.

Notably, what goes wrong in cases of *overplacement* is that our beliefs depart from reality. But how can we establish that people are committing a bias? After all, subjective beliefs do not admit easy verification, and one might really be better than the average. The obvious problem with the beliefs that guide our conception of our competence over that of others is, of course, that we cannot all be above average. Consequently, a significant proportion of us must be mistaken about our own relative insusceptibility to bias, suggesting that the relevant self-other asymmetry in fact reveals a tendency for *overplacement* in the accuracy of our judgments. In some cases, however, researchers have also tried to link these assessments with more objective measures. For instance, even people who have been hospitalized for accidents tend to believe that their driving skills are better than the average (e.g., Mckenna, Stanier and Lewis 1991).

6. Happiness research and goal-based rationality

It is also important to remember that other research projects besides that articulated by AR theorists have started to explore people's ability to achieve their goals rather than their ability to conform to rule-based rationality. In particular, in recent years decision scientists have started to directly study the contexts in which decisions succeed and fail to maximize happiness. They assess whether behaviour is adaptive, moving beyond coherence and consistency. For instance, Hsee and Hastie write that:

> For decades behavioural decision researchers have studied inconsistencies of choice. In recent years, however, researchers have studied directly when decisions fail to maximize happiness. (2006, 36)

Important methodological and theoretical issues arise from this psychological research (see, for instance, Haybron 2000; Alexandrova 2012; Angner 2013), but discussing them seems to fall beyond the scope of this section. Here I will just highlight two important points of which we should take note. First, it seems that this strand of research might be legitimately integrated into the AR project, because subjective well-being could be seen as one of the relevant goals: other than those that meet basic survival needs, or that adequately correspond to the external world, many decisions are motivated by the pursuit of subjective well being. The second relevant consideration here is that, *pace* AR theorists, most psychologists and behavioural researchers working in this field do not think that people make choices and decisions that maximize their happiness. So far, scholars have examined two general reasons for failure in this regard: (i) prediction biases and (ii) failures to follow predictions. As an illustration of the former, consider the so-called *impact bias*. People often

overestimate the impact (both intensity and duration) of an affective event. Junior faculty members typically overestimate the joy of getting tenure and the misery of being turned down (Gilbert et al. 1998). One cause of this *impact bias* is *focalism*—predictors pay too much attention to the central event and overlook context events that will moderate the central event's impact. For example, college football fans overpredicted the joy that they would experience in the days following the victory of their favoured team, because they failed to consider that the victory was only one of a myriad of events that would affect their future hedonic state (for more on this, see Kahneman 1997; Hsee and Hastie 2006; 2008).

7. Catching liars and "truth biases"

So far we have discussed a number of systematic biases mentioned in the psychological literature on judgment and decision-making that do not seem to have been identified or defined by relying on rule-based rationality. But now I want to stress that, if AR theorists want behavior and cognition to be assessed against goal-based rationality, other psychological evidence has to be considered that does not sit very well with their optimistic claims about people's rationality. In particular, I refer to psychological research on lie detection. This line of research is worthy of attention for two main reasons. First, the alleged ability to detect cheaters has attracted a great deal of attention from AR theorists, since they seem to consider this ability to be an important aspect of our "adaptive rationality" itself.[105] More precisely, according to

[105] Most discussions of these topics by AR theorists have focused on literature from experimental research on *social exchange:* interactions in which one party provides a benefit to the other conditional on the recipient's providing a benefit in return (e.g. Cosmides 1985; Cosmides and Tooby 1989; 1996).

them, as well as to several evolutionary behavioral scientists, humans have evolved a cognitive system that directs attention to information that could reveal cheaters. Second, as we saw in Chapter 3, critics of AR—such as Sterelny—have emphasized that heuristics cannot help us achieve goals such as catching liars.[106] While AR theorists often stress that we are remarkably good at detecting cheaters, here I want to draw attention to some overlooked findings from social psychology literature on people's inability to detect deceptive behavior accurately. In so doing, however, I do not aim to defend Sterelny's claim that simple heuristics cannot perform well in social environments. In fact, as I mentioned in Chapter 3, there are a number of problems with Sterelny's claim.[107] Instead, I merely claim that, in this particular context, it seems that people's performance does not support—and does not sit well with—the rather optimistic assessments of human rationality suggested by AR theorists.

So, what interests us here is that in several studies carried out by psychologist DePaulo and her co-workers, subjects were shown videotapes of people talking, where only the experimenters knew whether they were lying or telling the truth. The goal was to determine whether people could separate truth from lies (cf. DePaulo 1994).[108] The topics of these lies and truths varied widely. For example, sometimes the people on the tape talked about their feelings about other people they knew. At

[106] Specifically, as the reader will recall, Sterelny claims that 'it is no accident that the examples of such heuristics in action ignore interactions with other intelligent agents, especially competitive agents. For it is precisely in such situations that simple rules of thumb will go wrong […] Catching a ball is one problem; catching a liar is another' (Sterelny, 2003, 53).[106]

[107] Here one might actually want to emphasize also that what is unsuccessful behaviour for the subject trying to detect lies is at the same time an instance of successful behaviour from the perspective of the liar.

[108] This contemporary work has been shaped in important ways by Paul Ekman's original work on the detection of lies and emotions. For an overview of Ekman's research see, e.g., Ekman (1996).

other times, they gave their opinions on controversial issues; in still other studies, they talked to an artist about their preferences for various paintings, some of which were the artist's own work. When researchers showed people ("judges") these tapes, they asked them to decide, for each segment that they watched, whether they thought the person on the tape (the "speaker") was lying or telling the truth. Researchers also asked them to indicate, on a scale, just how deceptive or truthful the speaker seemed to be.

As it turned out, one consistent observation was that people's ability to detect deception is only slightly better than chance at just below 54%(Bond and DePaulo 2006). In particular, Kraut (1980) offered a statistical summary of results from ten such experiments. Finding a mean accuracy rate of 57%, Kraut concluded that the accuracy of human lie detection is low. In a summary of 39 studies published after 1980, Vrij (2000) replicated Kraut's findings, discovering that subjects achieve an average of 56.6% accuracy. These summaries have inspired a consensus: 'it is considered virtually axiomatic […] that individuals are at best inaccurate at deception detection' (Hubbell et al. 2001 115). Overall, such research suggests that people struggle to detect deception and the prevalence of truth-bias (Bond and DePaulo 2006), where the truth-bias is the tendency to believe other people independent of actual honesty (Levine et al. 1999).

8. Objections

The evidence mentioned above seems to suggest that AR theorists should be more careful in licensing optimistic claims about human rationality. Here I discuss three objections to my claim.

8.1 Sometimes we are adaptively irrational—so what?

The objector might claim that being optimistic about human rationality does not commit AR theorists to the claim that judgment and decision-making are *always* successful. In fact, it would be odd if AR theorists claimed that we are perfect decision-makers and that our reasoning was always successful. And, of course, they do not claim this. Thus, the objector can argue, evidence of unsuccessful reasoning and decision-making might not be seen as particularly problematic. This is an important point, but some remarks are in order. This consideration does not, in itself, justify AR theorists' failure to discuss the findings presented above, and it seems that these researchers have been overly selective in their discussion of empirical findings suggesting optimism about human rationality. Moreover, this omission is particularly important, since the biases discussed in this chapter do not seem to be just occasional failures of goal-based rationality, but rather systematic and widespread, and so worthy of special attention. In addition, it seems reasonable to ask these scholars to state more clearly how much unsuccessful reasoning and decision-making they can tolerate before dropping their optimistic claims about human rationality. In particular, while it is right to say that, of course, people's reasoning and decision-

making cannot be always successful, this should not become an excuse to overlook their failures and poor performance in some contexts and domains.

8.2 Biases are adaptive

One could also seek to resist my conclusions by arguing that such biases are in fact adaptive once measured in terms of conduciveness to well-being and fitness-maximization. In particular, during the past 15 years there has been an intense debate about the impact of unrealistic self-views on mental health (for a recent review of some literature on accurate cognition and mental health see, e.g., Bortolotti and Antrobus 2015). The dominant assertion in the debate is that overly positive views are actually beneficial for coping and psychological adjustment when people face extreme adversity (Taylor and Brown 1988), such as the aftermath of a civil war (Bonanno, Fiedl, Kovacevic and Kaltman 2002). There are, however, some problems with this objection.

A first consideration is that it is unclear whether we should accept this view as empirically well grounded. In fact, this research on the pragmatic benefits of flawed self-assessments is also controversial (Colvin and Block 1994). And people seem to agree that, if overrating one's self is advantageous, it is desirable only in moderate doses. In addition, the claim that misbeliefs can promote well-being and be evolutionarily adaptive is generally made with regard to only some of the biases discussed so far, namely *unrealistic optimism* and the *better than the average effect*,

and similar considerations do not seem to apply to other phenomena, such as the *planning fallacy* and *anchoring*.

Second, even if we assume that this research on adaptive misbeliefs is empirically well grounded, it is unclear what conclusions would follow from that. For instance, let us accept that, as McKay and Dennett (2009) suggest, some sets of misbeliefs are evolutionarily adaptive. The objector also needs to show that, in that case, it is fitness maximization, and not empirical accuracy, the relevant goal, and this seems to be problematic. Consider now the claim that, since some empirically inaccurate beliefs supposedly help manage negative emotions, these misbeliefs are psychologically adaptive. Notably, if empirical accuracy were at least one of the relevant goals of the cognizer, such misbeliefs would still count as suboptimal. Most plausibly, we should see an agent's success as resulting from a combination of goals and their relative importance, and it is not straightforward to deny that accuracy is at least one of the relevant goals here.

Third, and perhaps most importantly, by adopting this objection AR theorists would have to revise key tenets of their project. In the first place, they would have to acknowledge that their original diagnosis of the sins of MBR was inaccurate: while AR theorists have traditionally claimed that bias researchers have mistakenly assessed behavior against rule-based rationality, according to the objection presented above the problem is, instead, that bias researchers have assessed behavior against the wrong goals (empirical accuracy). More importantly, AR theorists have typically

assessed behavior against its empirical accuracy and maintained that empirical accuracy and agential success overlap. By stressing the adaptive value of empirically inaccurate beliefs, AR theorists would have to redefine and significantly amend core assumptions of their framework.

8.3 Biases are violations of rule-based rationality

Finally, the objector might argue that, after all, the abovementioned families of biases are at least suggestive of violations of rule-based rationality since these biases might involve an error over inductive or abductive reasoning. More precisely, while in the abovementioned cases subjects' behaviour was ultimately assessed against epistemic goals and people's desires and preferences, in committing these biases subjects have probably violated some norms of rule-based rationality. This might seem problematic for my argument. More specifically, it might be taken to show that biases are, after all, violations of rule-based rationality—where this is precisely the claim that I have been trying to attack.

The objection, however, does not seem to be particularly strong: even if we grant that such biases are really suggestive of reasoning that violates rule-based rationality, it does not seem that they have been individuated by relying on rule-based rationality or that they are constituted by the latter. Moreover, it is not clear that those biases really are suggestive of processes that violate rule-based rationality. This point becomes clearer when we consider a different and opposite reaction.

Let us move on to the second possible response. An objector might claim at this point that the families of biases considered in this chapter are instances of rational behaviour according to rule-based rationality, as people may arrive at such biases through standardly rational information processing. For instance, Dawes (1989) made such an argument with regard to the so-called *false consensus effect* (Ross, Greene and House 1977). This bias refers to a particular phenomenon: people who exhibit a particular view (performers) believe that this behaviour or view is more common overall than do people with different behaviour or view. Different explanations of the phenomenon have been given, but what is interesting for the present purpose is Dawes and Mulford's (1996) view that *false consensus* is in line with a Bayesian analysis that assumes a uniform prior distribution and one's own view as the only evidence. The idea is that the *false consensus effect* might be a consequence of a standard normatively appropriate strategy of generalizing from one indisputable datum—namely one's own response (Dawes 1989). On such view, rational information processing strategies might thus contribute to these biases: although demonstratively false (both groups cannot both be right), the belief is precisely what one would expect when people quite sensibly use what they know (their own belief or action) to inform what they do not know (the belief and actions of others). Note, however, that reinterpretations of other kinds of biases along similar lines are available (e.g., Benoit and Dubra 2011; Benoit, Dubra and Moore forthcoming; Lieder et al. 2012; Harris and Osman 2012). In light of such reinterpretations of these findings, one might think that we should consider these biases as instances of rational behaviour, or so the objector contends: we have biases

only in the sense that the observed beliefs do not match the actual distribution of outcomes, but these biases result from standardly rational processing.

Here I will offer a few considerations. First, it is controversial to what extent these models are psychologically realistic. For instance, it is hard to ignore the psychological evidence in the *heuristics-and-biases* tradition suggesting that individuals do not use Bayes' rule and, for that matter, may not even understand simple probability. There is a more important concern, however, which stresses that this kind of reply is not available to AR theorists. To see this, consider that, whilst AR theorists have launched their project as a radical normative departure from the perspective of rule-based rationality, by offering such reinterpretations they would be proposing a problematic resurrection of the normative perspective of rule-based rationality. What I want to suggest in this section, instead, is that the families of biases considered in this paper are compatible with processes that violate rule-based rationality as well as with those that conform to it.

9. Evolution and inaccurate reasoning

Finally, before I conclude this final chapter, there seems to be one important task left for us to accomplish. Specifically, the reader might now wonder how the outcome of the analysis carried out here relates to the considerations about the relationship between evolutionary thinking and psychological biases explored in Chapter 2. In particular, there I stressed that AR theorists had presented some considerations about the adaptive value of behavior that violates rule-based rationality in an attempt to

account for the evolution of some biases in MBR. But in this chapter I have shown that several biases discussed in MBR should not be characterized as mere violations of rule-based rationality, but rather as instances of mental contamination, empirically inaccurate beliefs and the like. So, one might still find it puzzling how such biases in MBR evolved. Here I seek to provide the reader with some thoughts and considerations on this topic.

First, whilst keeping in mind the limitations of the adaptationist account described in Chapters 2 and 6, it is important to highlight that in Chapter 2 I also presented some considerations regarding the adaptive value of empirically inaccurate beliefs. With regard to this point, I want to stress here that some models for the evolution of the instances of empirically inaccurate behavior described in this chapter have already been offered in the literature. For instance, Dennett himself, one of the authors generally associated with the main evolutionary argument discussed in Chapter 2 (EAR), seems to have moved away from his previous positions by challenging the presumption that true self-assessments are generally evolutionarily adaptive (McKay and Dennett 2009). In so doing, he discusses some clusters of misbeliefs, including unrealistic positive self-evaluations, exaggerated perception of personal control or mastery, and unrealistic optimism about the future. This is important: as we have seen in this chapter, there is a widespread tendency for most people to see themselves as better than most others in a range of aspects. Dennett suggests that manifestations of the *better than the average* effect might be not only pervasive but also evolutionarily adaptive. But other models for the evolution of such phenomena have recently been offered (e.g., Johnson and Fowler 2011).

It would be desirable to have evolutionary accounts for the other kinds of biased behavior reported here. Specifically, most debates on the evolution of rational behavior focus either on violations of axioms such as transitivity or on the evolution of true and false beliefs. But in acknowledging that there are other forms of "adaptive irrationality" and biased behavior besides violations of transitivity and empirically inaccurate beliefs, we thus call for more work on the evolution of these other sets of phenomena. Just to give an example here, consider that, as we have seen in this thesis, people are not very good at predicting what will make them happy. Notably, Buss suggested an evolutionary perspective on such forms of biased behavior, stressing that happiness is seen as 'a common goal toward which people strive, but for many […] remains frustratingly out of reach' (2000, 15). According to his view, there are plausible evolutionary reasons why humans do not excel at achieving happiness. In particular, Buss stresses that 'an evolutionary psychological perspective offers unique insights into some vexing barriers to achieving happiness and consequently into creating conditions for improving the quality of human life' (15) (see also Grinde 2002 and Ahuvia 2008, 502). I want to suggest that discussions about the evolution of human (ir)rationality should take into account the different phenomena and instances of biased behaviour that I have discussed in this chapter.

10. Conclusion

In this chapter I have articulated a reply to the challenge that AR theorists have raised to MBR. I have argued that, even if AR turned out to represent a plausible

normative perspective on rational behaviour, the challenge mounted by AR theorists seems to be less damaging than its advocates believe. First, there are reasons to think that, at least in several contexts, violations of rule-based rationality correspond to failures from the perspective of goal-based rationality as well. Second, I have called into question the claim that scholars in MBR have relied on rule-based rationality alone when individuating biases, and that what constitutes biases in MBR is that they are violations of rule-based rationality. I showed that, in fact, the category of bias refers to a rather heterogeneous class of phenomena and that the characterization of biases in terms of violations of rule-based rationality does not seem to fit several well-known families of biases. In light of this, I have argued that these families of biases do not seem to be vulnerable to the attacks on rule-based rationality mounted by AR theorists. In fact, many instances of biased behaviour look like instances of unsuccessful behaviour assessed against prudential and epistemic goals, and this does not seem to sit well with AR theorists' rather optimistic claims about human rationality. I concluded that either AR theorists explain clearly how these findings could be accommodated within an optimistic picture of human rationality, or they backpedal on their optimistic commitments.

8. Conclusion

It always takes longer than you expect,
even when you take into account Hofstadter's Law.
Hofstadter's Law

As we reach the end of our investigation, a recap of the main themes and achievements of this work will be useful. The following (probably containing some overlaps, repetitions and omissions) is a list of the main achievements and the core claims made in my dissertation.

This work started by observing that, while for several decades it has been commonplace in empirical debates on people's thinking and decision-making to think that humans are irrational, several research groups have recently started to question this view. In this work I have focused on the AR project and attempted to evaluate its scientific viability.

It became clear in our investigation that the question "Are humans rational?" should be broken down into two quite separate questions: "What does it mean to be rational?" and "Are humans rational?" With regard to the first question, AR theorists argue that people should not be assessed against norms of logic, probability theory or decision theory (or what I referred to in this work as rule-based rationality), but rather against the goals they entertain (or what we have dubbed goal-based rationality). With regard to the second question, AR theorists maintain that the conclusion that people are irrational is unsupported: people are often remarkably

successful once assessed against their goals and given the cognitive and external constraints imposed by the environment.

As I have tried to make clear, there are reasons why the challenge mounted by AR theorists is particularly important. Whilst other critics of MBR have focused mainly on methodological arguments related to the robustness of biases (cf. Chapter 1) or on evolutionary arguments against the very possibility of human irrationality (cf. Chapter 2), AR theorists have offered a more radical challenge: they have taken issue with some of the normative commitments of MBR, although they still share some of them—since they are interested in adaptive behaviour and cognition and in instrumental rationality.

This thesis sought to provide a qualified defence of such a program. On the one hand, I argued that there is room for a conceptual and methodological revolution in the study of rationality. Specifically, while it is commonly argued that to be rational means to reason according to formal principles of rationality (or rule-based rationality), I stressed the importance of assessing behaviour against the goals that people entertain (goal-based rationality). However, *contra* AR theorists, I also pointed to evidence suggesting that people are often remarkably unsuccessful at achieving epistemic and prudential goals, and argued that AR theorists have not made clear how these findings could be reconciled with their optimistic view of human rationality. Therefore, I concluded that, while AR theorists have indeed made significant progress in the "rationality debate", they have hitherto failed to provide compelling evidence or arguments in support of their most ambitious theses.

Now, I am aware that one should not derive momentous implications from literally interpreting isolated statements, and that some exaggerations on the side of AR may be explained in light of a need to obtain public attention and funding. Even so, one would expect AR theorists to advance much more measured claims in the future, for their propensity to overstate their own achievements has generated a lot of unnecessary confusion in the literature, leading many scholars interested in the "rationality debate" to be needlessly sceptical about the prospects of AR.

At this point, it is worth taking a closer look at the structure of my general strategy. The main claims that I have just summarized have been reached through a careful examination of a number of arguments, and the core of my defence of AR theorists' attack on reliance on rule-based rationality was offered in Chapters 3, 4 and 6. First, I noted that emphasis on the importance of adaptive behaviour and instrumental views on rationality is extremely popular in the psychological literature on judgement and decision-making and, importantly, generally advocated by AR theorists and MBR researchers. Second, I showed that in several contexts and domains behaviour that violates rule-based rationality can be adaptive and successful. This outcome raises a non-trivial problem for MBR researchers: if what ultimately justifies rule-based rationality is the promise of success, it seems hard to count as irrational violations of these norms that lead to successful and adaptive behaviour. Later, in Chapter 6, I looked at research on individual differences in decision-making and examined whether people who score higher in cognitive ability tests are more likely to achieve desired life outcomes because they are less prone to violating rule-based rationality. *Prima facie*, such research could represent an

interesting large-scale project in judgement and decision-making and show that following rule-based rationality is, after all, conducive to successful behaviour. However, I concluded that, as the matter stands, we are unable to draw causal conclusions about whether conformity to rule-based rationality is conducive to successful behaviour.

My arguments concerning AR theorists' case for optimism about human rationality were presented in Chapter 7. Specifically, in Chapter 7 I provided reasons to resist AR theorists' optimism about human rationality. In the main, my strategy consisted in showing that a great deal of research on biases and irrational behaviour is in fact unscathed by AR theorists' arguments: numerous families of biases cannot be described as mere violations of rule-based rationality, and in several cases it seems that behaviour has been assessed against epistemic and prudential goals after all. As it turns out, this evidence does not seem to sit well with the conclusions about human rationality drawn by AR theorists. In particular, AR theorists have not made how, in light of this evidence, they could still hold their optimistic picture of human rationality.

I hope that my journey into AR will constitute a step forward for the study of (ir)rational behaviour. Admittedly, several issues still remain unsolved, and for a number of questions that have been answered in this work, a whole new set of questions has arisen. But a further goal of this thesis was precisely to draw attention to some previously overlooked issues. In particular, I sought to prompt AR theorists to build their case on more solid conceptual foundations. I argued that AR theorists

(but scholars in judgement and decision-making more generally) have not articulated in enough detail the notions of "goal" and "goal attainment". As I tried to show in Chapter 4, AR theorists, as well as other scholars engaged in the "rationality debate", tend to appeal to such notions, but also tend to switch back and forth between different understandings of them. Whilst I illustrated some possible ways of articulating the relevant perspective of goal-based rationality, I also emphasized that further refinement and more work are needed if AR theorists wish to present their framework as a convincing alternative normative framework for the assessment of rational behaviour and cognition.

Finally, the reader might also draw from this work a number of more general lessons. For instance, a general consideration made here was about previous discussions of the evolution of rational and irrational behaviour that have been offered in the literature. I argued that scholars interested in the "rationality debate" had not paid enough attention to the actual details of results suggesting people's irrationality given by MBR scholars. One way in which these discussions have generally mischaracterized the psychological phenomena in question is by overlooking the existence of individual differences in judgements and decision-making. In light of the discussion offered in Chapter 6, however, it should be clear that it is not sufficient to provide an evolutionary model for the fact that people are biased with regard to some particular task or behaviour. What needs to be explained is that, in most cases, some people are biased and others are not. But there is also another way in which discussions of the evolution of irrational behaviour have gone astray. Specifically, commentators have typically failed to appreciate that the range

of phenomena classed as biases is rather heterogeneous. In particular, researchers interested in offering accounts of the evolution of biases have typically either focused on a set of empirically inaccurate beliefs or on violations of some principles of rule-based rationality, quite often violations of transitivity. In so doing, however, they have focused on a rather narrow class of biases. In light of the considerations offered in Chapter 7, it is clear that, if researchers are interested in accounting for the evolution of biased behaviour, a broader range of phenomena must be taken into account and accounted for.

References

Aarts, H., and Dijksterhuis, A. (2000). Habits as knowledge structures: automaticity in goal-directed behavior. *Journal of Personality and Social Psychology*, 78: 53-63.

Adam, M. B., and Reyna, V. F. (2005). Coherence and correspondence criteria for rationality: Experts' estimation of risks of sexually transmitted infections. *Journal of Behavioural Decision Making*, 18: 169-186.

Ahuvia, A. (2008). If money doesn't make us happy, why do we act as if it does? *Journal of Economic Psychology,* 29: 491-507.

Alcock, J. (2005). *Animal behaviour: An evolutionary approach* (8th ed.). Sinauer Associates.

Alexandrova, A. (2012). Well-being as an object of science. *Philosophy of Science*, *79*: 678-689.

Alicke M. D. (1985) Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, 49: 1621–1630.

Anand, P. (1987). Are the preference axioms really rational? *Theory and Decision*, 23: 189-214.

Anderson, J. R. (1990). *The Adaptive Character Of Thought*. Lawrence Erlbaum.

Angner, E. (2013). Is it possible to measure happiness? *European Journal for Philosophy of Science*, 3: 221-240.

Ariely, D. (2009). *Predictably irrational*. Harper Collins.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, 110: 486-498.

Arkes, H. R., and Ayton, P. (1999). The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin*, 125: 591–600.

Arló-Costa, H., and Pedersen, A. P. (2011). Bounded rationality: Models for some fast and frugal heuristics. In A. Gupta, J. F. A. K. van Benthem, and E. Pacuit (Eds.), *Games, norms and reasons: Logic at the crossroads. A tribute to Rohit Parikh on the occasion of his 70th birthday*. Springer.

———————————————. (2012). Fast and frugal heuristics, rationality, and the limits of naturalism, *Synthese,* 190: 831-850.

Arnau, E., Ayala, S., and Sturm, T. (2014). Cognitive externalism meets bounded rationality. *Philosophical Psychology*, 27: 50-64.

Ayton, P., and Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & cognition*, 32: 1369-1378.

Azar, O. H. (2008). The effect of relative thinking on firm strategy and market outcomes: A location differentiation model with endogenous transportation costs. *Journal of Economic Psychology,* 29: 684-697.

——————. (2011). Do consumers make too much effort to save on cheap items and too little to save on expensive items? Experimental results and implications for business strategy. *American Behavioral Scientist,* 55: 1077-1098.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer (Ed.), *The automaticity of everyday life: Advances in social cognition*. Erlbaum.

Bargh, J. A., and Gollwitzer, P. M. (1994). Environmental control of goal-directed action: automatic and strategic contingencies between situations and behavior. In W. D. Spaulding (Ed.), *Nebraska Symposium on Motivation*. University of Nevada Press.

Baron, J. (1985). *Rationality and intelligence*. Cambridge University Press.

———. (2000). Normative and prescriptive implications of individual differences. *Behavioral and Brain Sciences*, 23: 668-669.

———. (2004). Normative models of judgment and decision-making. In D. J. Koehler and N. Harvey (Eds.), *The Blackwell handbook of judgment and decision-making*. Blackwell.

———. (2012). The point of normative models in judgment and decision-making. *Frontiers in psychology*, 3.

Batty, G. D., Deary, I. J., and Macintyre, S. (2007). Childhood IQ in relation to risk factors for premature mortality in middle-aged persons: The Aberdeen Children of the 1950s Study. *Journal of Epidemiology and Community Health*, 61: 241–247.

Batty, G. D., Deary, I. J., Schoon, I., and Gale, C. R. (2007). Mental ability across childhood in relation to risk factors for premature mortality in adult life: the 1970 British Cohort Study. *Journal of Epidemiology and Community Health*, 61: 997-1003.

Baumeister, R. F., Vohs, K. D., and Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, 16: 351-355.

Bazerman, M. H., and Banaji, M. R. (2004). The social psychology of ordinary ethical failures. *Social Justice Research*, 17: 111–115.

Beauducel, A., and Kersting, M. (2002). Fluid and crystallized intelligence and the Berlin Model of Intelligence Structure (BIS). *European Journal of Psychological Assessment*, 18: 97-112.

Becker, G. S., and Murphy, K. M. (1988). A Theory of Rational Addiction. *Journal of Political Economy,* 96: 675–700.

Behavioural Insights Team. (2010). *Applying Behavioural Insights to Health*, Cabinet Office.

Belsky, J., Steinberg, L., and Draper, P. (1991). Childhood experience, interpersonal development, and reproductive strategy: An evolutionary theory of socialization. *Child Development*, 62: 647-670.

Benoît, J. P., and Dubra, J. (2011). Apparent overconfidence. *Econometrica,* 79: 1591-1625.

Benoît, J. P., Dubra, J., and Moore, D. A. (*Forthcoming*). Does the better than the average effect show that people are overconfident? Two experiments. *Journal of the European Economic Association*.

Berkeley, D., and Humphreys, P. (1982). Structuring decision problems and the bias heuristic. *Acta Psychologica*, 50: 201-52.

Bermúdez, J. L. (2003). *Thinking without words*. Oxford University Press.

————————. (2009). *Decision Theory and Rationality*. Oxford University Press.

Binmore, K. (1999). Why experiment in economics? *Economic Journal*, 109: 16-24.

Blavatskyy, P., and Pogrebna, G. (2010). Endowment effects? "Even" with half a million on the table! *Theory and decision*, 68: 173-192.

Bonanno, G. A., Field, N. P., Kovacevic, A., and Kaltman S. (2002) Self-enhancement as a buffer against extreme adversity: Civil war in Bosnia and traumatic loss in the United States. *Personality and Social Psychology Bulletin*, 28: 184–196.

Bond, C. F., and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and social psychology Review*, 10: 214-234.

Bonini, N., Tentori, K., and Osherson, D. (2004). A different conjunction fallacy. *Mind & Language*, 19: 199–210.

Bookstaber, R., and Langsam, J. (1985). On the optimality of coarse behaviour. *Journal of theoretical biology*, 116: 161-193.

Bordley, R., and Hazen, G. B. (1991). SSB and weighted linear utility as expected utility with suspicion. *Management Science*, 37: 396-408.

Borges, B., Goldstein, D., Ortmann, A., and Gigerenzer, G. (1999) Can ignorance beat the stock market? In Gigerenzer G, Todd P, and the ABC Research Group (Eds.) *Simple heuristics that make us smart*. Oxford University Press.

Bortolotti, L. (2005). Intentionality without rationality. In *Proceedings of the Aristotelian Society,* CV, 385–392.

————————. (2011). Does reflection lead to wise choices? *Philosophical Explorations*, 14: 297-313.

————————. (2014). *Irrationality*. John Wiley & Sons.

Bortolotti, L., and Antrobus, M. (2015). Costs and benefits of realism and optimism. *Current Opinions in Psychiatry*, 28: 194–198.

Botterill, G., and Carruthers, P. (1999). *The philosophy of psychology*. Cambridge University Press.

Boudry, M., and Vlerick, M. (2014). Natural selection does care about truth. *International Studies in the Philosophy of Science*, 28: 65-77.

Boudry, M., Vlerick, M., and McKay, R. (2015). Can evolution get us off the hook? Evaluating the ecological defense of human rationality. *Consciousness and cognition*, 33: 525-535.

Bourgeois-Gironde, S., and Giraud, R. (2009). Framing effects as violations of extensionality. *Theory and Decision*, 67: 385-404.

Boyd, M. (2001). On ignorance, intuition and investing: a bear market test of the recognition heuristic. *Journal of Psychology and Financial Markets*, 2: 150–56.

Bowers, J. S. and Davis, C. J. (2012a). Bayesian Just-so Stories in Psychology and Neuroscience. *Psychological Bulletin*, 138: 389-414.

———————————. (2012b). Is that What Bayesians Believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin*, 138: 423-436.

Brainerd, C. J., and Reyna, V. F. (2005). *The science of false memory*. Oxford University Press.

Brännmark, J., and Sahlin, N. E. (2010). Ethical theory and the philosophy of risk: first thoughts. *Journal of Risk Research*, 13: 149-161.

Brighton, H., and Todd, P. M. (2009). Situating rationality: Ecologically rational decision making with simple heuristics. In P. Robbins and M. Aydede (Eds.), *Cambridge handbook of situated cognition*. Cambridge University Press.

Bröder. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29: 611–625.

Broome, J. (1999). *Ethics out of Economics*. Cambridge University Press.

Brown, H. I. (1989). *Rationality*. Routledge.

Brownstein, M. (2015). Implicit bias. *Stanford Encyclopedia of Philosophy*.

Brunswik, E. (1944). *Outline of history of psychology*. University of California Press.

————————. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review,* 62: 193-217.

————————. (1957). Scope and aspects of the cognitive problem. In H. Gruber, K. R. Hammond, and R. Jessor (Eds.), *Contemporary approaches to cognition*. Harvard University Press.

Buehler, R., Griffin, D., and Ross, M. (2002) Inside the planning fallacy: The causes and consequences of optimistic time predictions. In T. D. Gilovich, D. W. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Buller, D. J. (2005). *Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature*. MIT Press.

Bullock, S., and Todd, P. M. (1999). Made to measure: ecological rationality in structured environments. *Minds and Machines*, 9: 497-541.

Burns, B. D. (2001). The hot hand in basketball: Fallacy or adaptive thinking. In *Proceedings of the twenty-third annual meeting of the cognitive science society*, 152-157.

————————. (2004). Heuristics as beliefs and as behaviors: The adaptiveness of the "hot hand". *Cognitive Psychology*, 48: 295-331.

Burns, B. D., and Corpus, B. (2004). Randomness and inductions from streaks: "Gambler's fallacy" versus "hot hand". *Psychonomic Bulletin & Review*, 11: 179-184.

Buss, D. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioural and Brain Sciences*, 12: 1–49.

Butterfill, S. A., and Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, 28: 606-637.

Byrne S., and Whiten, A. (1988) *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.

Calvin, C. M., Deary, I. J., Fenton, C., Roberts, B. A., Der, G., Leckenby, N., and Batty, G. D. (2010). Intelligence in youth and all-cause-mortality: systematic review with meta-analysis. *International journal of epidemiology*, 40: 626-644.

Camerer, C. F. (2000). Prospect theory in the wild: evidence from the field. In Kahneman and A. Tversky (Eds.), *Choices, Values and Frames*. Cambridge University Press.

————————. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F., and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, 19: 7-42.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54: 297-312.

Camperio Ciani, A., Veronese, V., Capiluppi, and C., Sartori, G. (2007). The adaptive values of personality differences revealed by small island population dynamics. *European Journal of Personality*, 21: 3-22.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford University Press.

————————. (2007). Simple heuristics meet massive modularity. In P. Carruthers, S. Laurence and S. P. Stich (Eds.), *The Innate Mind: Culture and Cognition*. Oxford University Press.

Carruthers, P. (2012). The fragmentation of reasoning. In *P. Quintanilla, C. Mantilla, and P. Cépeda (eds.), Cognición Social y Lenguaje: La intersubjetividad en la evolución de la especie y en el desarrollo del niño,* Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú, 2014.

Casscells, W., Schoenberger, A., and Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299: 999-1001.

Castelfranchi, C. (2014). Intentions in the Light of Goals. *Topoi*, 33: 103-116.

Charness, G., Karni, E., and Levin, D. (2010). On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior,* 68: 551–556.

Charness, G., and Sutter, M. (2012). Groups make better self-interested decisions. *The Journal of Economic Perspectives*, 26: 157-176.

Chase, V. M., Hertwig, R., and Gigerenzer, G. (1998). Visions of Rationality. *Trends in Cognitive Sciences*, 2: 206–14.

Chater, N., and Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, 122: 93-131.

Chater, N., M. Oaksford, R. Nakisa and M. Redington. (2003). Fast, Frugal, and Rational: How Rational Norms Explain Human Behaviour. *Organizational Behaviour and Human Decision Processes*, 90: 63-86.

Cherniak, C. (1984). Computational complexity and the universal acceptance of logic. *Journal of Philosophy*, 81: 739-758.

Cherniak C. (1986). *Minimal rationality*. MIT Press.

Chernoff, H. (1954). Rational selection of decision functions. *Econometrica,* 22: 423- 443.

Chirimuuta, M., and Gold, I. (2009). The embedded neuron, the enactive field? In J. Bickle (Ed.), *The Oxford Handbook of Philosophy and Neuroscience*. Oxford University Press.

Chow, S. J. (forthcoming). Many Meanings of 'Heuristic'. *The British Journal for the Philosophy of Science*.

Chugh, D., Bazerman, M. H., and Banaji, M. R. (2005). Bounded ethicality as a psychological barrier to recognizing conflict of interest. In D. A. Moore, D. M. Cain, G. Lowenstein, and M. H. Bazerman (Eds.), *Conflict of interest: Challenges and solutions in business, law, medicine, and policy*. Cambridge University Press.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. MIT Press.

———. (2001). Reasons, Robots and the Extended Mind. *Mind and Language*, 16: 121-145.

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis*, 58: 7-19.

Cohen, L. J. (1981). Can human rationality be experimentally demonstrated? *Behavioural and Brain Sciences*, 4: 317–70.

Cokely, E. T., and Feltz, A. (2014). Expert intuition. In L. M. Osbeck, and B. S. Held (Eds.), *Rational intuition: Philosophical roots, scientific investigations*. Cambridge University Press.

Colvin, C. R., and Block, J. (1994). Do positive illusions foster mental health? An examination of the Taylor and Brown formulation. *Psychological Bulletin*, 116: 3-20.

Conway, M. A., and Loveday, C. (2015). Remembering, imagining, false memories & personal meanings. *Consciousness and Cognition*, 35: 574-581.

Cooper W (1989) How evolutionary biology challenges the classical theory of rational choice. *Biology and Philosophy*, 4: 457-481.

————. (2003). *The evolution of reason: Logic as a branch of biology*. Cambridge University Press.

Cooper, A. C., Woo, C. Y., and Dunkelberg, W. C. (1988) Entrepreneurs' perceived chances for success. *Journal of Business Venturing*, 3: 97-108.

Cosmides, L. (1985). *Deduction or Darwinian algorithms? An explanation of the "elusive" content effect on the Wason selection task*. Doctoral dissertation, Harvard University.

————. (1989). The logic of social exchange: Has natural selection shaped how humans reason? *Cognition*, 31: 187-276.

L. Cosmides, and Tooby, J. (1989). Evolutionary psychology and the generation of culture, Part II. Case study: A computational theory of social exchange. *Ethology and Sociobiology*, 10: 51–97.

————————————. (1992). Cognitive adaptations for social exchange. In H. Barkow, L. Cosmides and J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press.

————————————. (1994). Better than Rational: Evolutionary Psychology and the Invisible Hand. *American Economic Review*, 84: 327-332.

———————————. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58: 1-73.

Costa-Gomes, M. A., and Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: An experimental study. *The American economic review*, 96: 1737-1768.

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Oxford University Press.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, 25: 1-17.

Dawes, R. M., and Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behaviour and Human Decision Processes*, 65: 201-211.

Deary, I. (2008). Why do intelligent people live longer? *Nature*, 456: 175-176.

———. (2012). Looking for "System Integrity" in Cognitive Epidemiology. *Gerontology*, 58: 545-553.

Deary, I. J., Batty, G. D., Pattie, A., and Gale, C. R. (2008). More Intelligent, More Dependable Children Live Longer: A 55-Year Longitudinal Study of a Representative Sample of the Scottish Nation. *Psychological Science*, 19: 874-880.

Deary, I. J., Gale, C. R., Stewart, M. C., Fowkes, F. G. R., Murray, G. D., Batty, G. D., and Price, J. F. (2009). Intelligence and persisting with medication for two years: Analysis in a randomized controlled trial. *Intelligence*, 37: 607-612.

Deary, I. J., Weiss, A., and Batty, G. D. (2010). Intelligence and personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities. *Psychological Science in the Public Interest*, 11: 53-79.

Dennett, D. (1987). *The intentional stance*. MIT Press.

DePaulo, B. M. (1994). Spotting lies: Can humans learn to do better? *Current Directions in Psychological Science*, 3: 83-86.

Devetag, G., Ceccacci, F., and De Salvo, P. (2013). Do Reputation Concerns Make Behavioral Biases Disappear? The Conjunction Fallacy on Facebook and Mechanical Turk. *The Conjunction Fallacy on Facebook and Mechanical Turk*.

Dhami, M., Hoffrage, K., and Hertwig, R. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bullettin,* 130: 959-988.

Díez, J., and Lorenzano, P. (2013). Who got what wrong? Fodor and Piattelli on Darwin: Guiding principles and explanatory models in natural selection. *Erkenntnis*, 78: 1143-1175.

Dijksterhuis, A, Maarten, W., Nordgren, L., and van Baaren, R. (2006). On making the right choice: the deliberation-without-attention effect. *Science*, 311: 1005–1007.

DiMatteo, M. R. (2004). Variations in patients' adherence to medical recommendations: A quantitative review of 50 years of research. *Medical Care*, 42: 200–209.

Dolan, P., Hallsworth, M., Halpern, D., King, D., and Vlaev, I. (2010). *MINDSPACE: influencing behavior through public policy*. Institute for Government.

Dukas, R. (Ed.). (1998). *Cognitive ecology: The evolutionary ecology of information processing and decision making*. University of Chicago Press.

Dunning, D., Heath, C., and Suls, J. M. (2004). Flawed self-assessment implications for health, education, and the workplace. *Psychological science in the public interest*, 5: 69-106.

Edwards, W. (1954). The Theory of Decision Making. *Psychological bulletin,* 51: 380-417.

————————. (1966). *Nonconservative information processing systems*. University of Michigan Press.

————————. (1983). Human cognitive capabilities, representativeness, and ground rules for research. In P. Humphreys, O. Svenson, and A. Vari (Eds.), *Analysing and aiding decision processes (vol. 18)*. North-Holland Publishing Company.

Ekman, P. (1996). Why don't we catch liars? *Social research*, 63: 801-817.

Elqayam, S. (2012). Grounded rationality: Descriptivism in epistemic context. *Synthese*, 189: 39-49.

Elqayam, S., and Evans, J. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences* 34: 251-252.

Erceg, N., and Galić, Z. (2014). Overconfidence bias and conjunction fallacy in predicting outcomes of football matches. *Journal of Economic Psychology*, 42: 52-62.

Evans, J. S. B. T., and Over, D. E. (1996). *Rationality and Reasoning*. Psychology Press.

Evans, J. S. B., and Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, 8: 223-241.

Fantino, E., and Stolarz-Fantino, S. (2005). Decision-making: Context matters. *Behavioural Processes*, 69: 165–171.

Fawcett, T. W., Fallenstein, B., Higginson, A. D., Houston, A. I., Mallpress, D. E., and McNamara, J. M. (2014). The evolution of decision rules in complex environments. *Trends in Cognitive Sciences*, 18: 153-161.

Fehr, E., and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425: 785–791.

Fernández, J. (2014). What are the benefits of memory distortion? *Consciousness and Cognition,* 33: 536-547.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, 50: 123–129.

Finucane, M. L., Alhakami, A., Slovic, P., and Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, 13: 1-17.

Fischbacher, U., Gachter, S., and Fehr, E. (2001) Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters*, 71: 397–404.

Fishburn, P. C. (1973). *The theory of social choice*. Princeton University Press.

———————. (1991). Nontransitive preferences in decision theory. *Journal of Risk and Uncertainty*, 4: 113-134.

Fisk, J. E., and R. Slattery. (2005). Reasoning about conjunctive probabilistic concepts in childhood. *Canadian Journal of Experimental Psychology*, 59: 168–178.

Fodor, J. (1981). *Representations*. MIT Press.

————. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and brain sciences*, 3: 63-73.

Fodor, J., and Piattelli-Palmarini, M. (2010). *What Darwin Got Wrong*. Farrar, Straus and Giroux.

Frankfurt, H. G. (1982). The Importance of What We Care About. *Synthese,* 53: 257-272.

Frederick, S. (2005). Cognitive reflection and decision-making. *Journal of Economic perspectives*, 19: 25-42.

Fulda, F. (2015). A mechanistic framework for Darwinism or why Fodor's objection fails. *Synthese*, 192: 163-183.

Futuyma, D. J. (2010). Two critics without a clue. *Science*, 328: 692-693.

Gächter, S., Orzen, H., Renner, E., and Starmer, C. (2009). Are experimental economists prone to framing effects? A natural field experiment. *Journal of Economic Behavior & Organization*, 70: 443-446.

Gale, C. R., Deary, I. J., Boyle, S. H., Barefoot, J., Mortensen, L. H., and Batty, G. D. (2008). Cognitive ability in early adulthood and risk of 5 specific psychiatric disorders in middle age: the Vietnam experience study. *Archives of General Psychiatry*, 65: 1410-1418.

Gangestad, S. W., and Simpson, J. A. (1990). Toward an evolutionary history of female sociosexual variation. *Journal of Personality*, 58: 69-96.

Garcia J, McGowan, B. K., and Green, K. F. (1972) Biological constraints on conditioning. In Black A. H., and Prokasy, W. F. (Eds.), *Classical conditioning II: current research and theory*. Erlbaum.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. Basic Books.

Gibson, G. (2005). The origins of stability: Review of Robustness and Evolvability in Living Systems by Andreas Wagner. *Science*, 310: 237-237.

Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *The American Journal of Psychology*, 97: 185-195.

———————. (1991). How to make cognitive illusions disappear. *Review of Social Psychology,* 45: 83–115.

———————. (1993). The bounded rationality of probabilistic mental models. In K. I. Manktelow and D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives*. Routledge.

———————. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology. In G. Wright and P. Ayton (Eds.), *Subjective Probability*. John Wiley & Sons.

———————. (1995). The taming of content; some thoughts about domains and modules. *Thinking and Reasoning*, 1: 324-335.

———————. (1996a). The psychology of good judgment Frequency formats and simple algorithms. *Medical Decision Making*, 16: 273-280.

———————. (1996b). On narrow norms and vague heuristics: a reply to Kahneman and Tversky (1996). *Psychological Review,* 103: 592–596.

———————. (1997). The modularity of social intelligence. In A. Whiten and R. Byrne (Eds.), *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge University Press.

———————. (1998). Ecological intelligence: an adaptation for frequencies. In D. Cummins and C. Allen (Eds.), *The Evolution of Mind*. Oxford University Press.

———————. (2000). *Adaptive Thinking: Rationality in the Real World*. Oxford University Press.

———————. (2001). Ideas in exile: The struggles of an upright man. In K. R. Hammond and T. R. Stewart (Eds.), *The essential Brunswik: Beginning, explications, applications*. Oxford University Press.

———————. (2002). *Reckoning with risk: learning to live with uncertainty*. Penguin Books.

———————. (2004). The irrationality paradox. *Behavioral and Brain Sciences*, 27: 336-338.

———————. (2007). *Gut Feelings: The Intelligence of the Unconscious*. Penguin Books.

———————. (2008). Moral Intuition = Fast and Frugal Heuristics? In Sinnott-Armstrong, W. (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*. MIT Press.

———————. (2008). Why heuristics work? *Perspectives on psychological science*, 3: 20-29.

Gigerenzer, G., and Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1: 107-143.

Gigerenzer G, and Goldstein, D. G. (1996) Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103: 650–669.

Gigerenzer, G., and Gray, J. M. (2011). *Better doctors, better patients, better decisions: Envisioning health care 2020*. MIT Press.

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102: 684-704.

Gigerenzer, G., Hoffrage, U., and Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review,* 98: 506-28.

Gigerenzer, G., and K. Hug. (1992). Domain-specific reasoning: social contracts, cheating, and perspective Change. *Cognition,* 43: 127–171.

Gigerenzer, G., and Gaissmaier, W. (2011) Heuristic Decision Making. *Annual Review of Psychology*, 62: 451–482.

Gigerenzer, G., and Sturm, T. (2012). How (far) can rationality be naturalized? *Synthese*, 187: 243-268.

Gigerenzer G, Todd, P. M., and the ABC Research Group. (1999) *Simple heuristics that make us smart*. Oxford University Press.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., and Wheatley, T. P. (1998). Immune neglect: a source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75: 617-638.

Gilboa, I. (2010). *Rational choice*. MIT Press.

Gilboa I., Postlewaite, A., and Schmeidler, D. (2012) Rationality of belief or: why savage's axioms are neither necessary nor sufficient for rationality. *Synthese,* 187: 11-31.

Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics & Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17: 295–314.

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of question form and information structure. *Cognition*, 78: 247-276.

———————————. (2008). Children's understanding of posterior probability. *Cognition*, 106: 325-344.

Glöckner, A., Hilbig, B. E., and Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition*, 133: 641-666.

Goeree, J. K., and Holt, C. A. (2002). Private Costs and Public Benefits: Unravelling the Effects of Altruism and Noisy Behavior. *Journal of Public Economics*. 83: 257–278.

Goldman, A., and Mason, K. (2007). Simulation. In P. Thagard (Ed.), *Handbook of philosophy of psychology and cognitive science*. Elsevier.

Goldstein, D. G., and Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological review*, 109: 75-90.

Goldstein, D. G., and Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, 9: 1-14.

Good I. J. (1983). *Good thinking: The foundations of probability and its applications*. University of Minnesota Press.

Goodman, N. (1965). *Fact, fiction, and forecast*. Bobbs-Merrill.

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. *Intelligence*, 24: 79–132.

———————————. (2004). Intelligence: Is it the epidemiologists' elusive ''fundamental cause'' of social class inequalities in health? *Journal of Personality and Social Psychology,* 86: 174-199.

Gottfredson, L. S., and Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science*, 13: 1-4.

Gowda, M. V. R., and Fox, J. C. (2002). *Judgments, decisions, and public policy*. Cambridge University Press.

Graham, P. A. (2011). 'Ought' and Ability. *Philosophical Review* 120: 337-382.

Grether, D. M., and Plott, C. R. (1979). Economic theory of choice and the preference reversals phenomenon. *American Economic Review*, 69: 623-638.

Griffiths, T. L., Chater, N., Norris, D., and Pouget, A. (2012). How the Bayesians Got their Beliefs (and what those beliefs actually are): Comments on Bower and Davis (2012). *Psychological Bulletin*, 138: 415-22.

Grinde, B. (2002). Happiness in the perspective of evolutionary psychology. *Journal of Happiness Studies,* 3: 331-54.

Grüne-Yanoff, T. (2007). Bounded rationality. *Philosophy Compass*, 2: 534-563.

———————. (2012). Paradoxes of rational choice theory. In Roeser, S., Hillerbrand, R., Sandin, P., and Peterson, M. (Eds.), *Handbook of Risk Theory*. Springer.

Grüne-Yanoff, T., and Hertwig, R. (Forthcoming). Nudge Versus Boost: How Coherent are Policy and Theory? *Minds and Machines*.

Guala, F. (2005). *The methodology of experimental economics*. Cambridge University Press.

Hahn, U., and Harris, A. J. (2014). What does it mean to be biased: motivated reasoning and rationality. *PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL*. 61: 41-102.

Hahn, A., Judd, C. M., Hirsh, H. K., and Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143: 1369-1392.

Hammond K. R. (1996) *Human judgment and social policy*. Oxford University Press.

———————. (2007) *Beyond rationality*. Oxford University Press.

Hammond, K. R., and Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford University Press.

Hands, D. W. (2014). Normative ecological rationality: normative rationality in the fast-and-frugal-heuristics research program. *Journal of Economic Methodology,* 21: 396-410.

Hanoch, Y. (2002). "Neither an angel nor an ant": Emotion as an aid to bounded rationality. *Journal of Economic Psychology*, 23: 1-25.

Hansson, S. O., and Grüne-Yanoff, T. (2006). Preferences. *Stanford Encyclopedia of Philosophy*.

Harman, G. (1986). *Change in view: Principles of reasoning*. MIT Press.

Harman, G. (1995). Rationality. In E. E. Smith and D. N. Osherson (Eds.), *Thinking: An Invitation to Cognitive Science, Volume 3*. MIT Press.

Harris, A. J., and Osman, M. (2012). The illusion of control: A Bayesian perspective. *Synthese*, 189: 29-38.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature,* 42: 1009- 1055.

Hart, S. A., Petrill, S. A., Thompson, L. A., and Plomin, R. (2009). The ABCs of math: A genetic analysis of mathematics and its links with reading ability and general cognitive ability. *Journal of Educational Psychology*, 101: 388-402.

Hart, C. L., Taylor, M. D., Davey Smith, G., Whalley, L. J., Starr, J. M., Hole, D. J., Wilson, V., and Deary, I. J. (2003). Childhood IQ, social class, deprivation and their relationships with mortality and morbidity risk in later life. *Psychosomatic Medicine*, 65: 877–883.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., and Moore, T. (2009). Adaptive rationality: an evolutionary perspective on cognitive bias. *Social Cognition*, 27: 732-762.

Haselton M. G., Buss D. M. (2000) Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, 78: 81–91.

Hastie, R., and R. Dawes. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Hájek, A. (1996). "Mises redux"—redux: Fifteen arguments against finite frequentism. *Erkenntnis,* 45: 209-227.

————. (2009). Fifteen arguments against hypothetical frequentism. *Erkenntnis*, 70: 211-235.

Hausman, D. M. (2011). *Preference, value, choice, and welfare*. Cambridge University Press.

Hausman, D. M., and Welch, B. (2010). Debate: To Nudge or Not to Nudge. *Journal of Political Philosophy*, 18: 123-136.

Haworth, C. M. A., Wright, M. J., Luciano, M., Martin, N. G., De Geus, E. J. C., Van Beijsterveldt, C. E. M. and Plomin, R. (2009). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry*, 15: 1112-1120.

Haybron, D. M. (2000). Two philosophical problems in the study of happiness. *Journal of Happiness Studies*, 1: 207-225.

Hazlett, A. (2013). *A luxury of the understanding: on the value of true belief*. Oxford University Press.

Hertwig, R., and Chase, V. M. (1998). Many reasons or just one: how response mode affects reasoning in the conjunction problem. *Thinking and Reasoning*, 4: 319–352.

Hertwig, R., and Gigerenzer, G. (1999). The conjunction fallacy revisited: how intelligent inferences look like reasoning errors. *Journal of Behavioural Decision Making*, 12: 275-305.

Hertwig, R., and Hoffrage, U. (2013). *Simple heuristics in a social world*. Oxford University Press.

Hertwig, R., and Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24: 383-403.

Hertwig, R., Pachur, T., and Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31: 621–642.

Hertwig, R., and Volz, K. G. (2013). Abnormality, rationality, and sanity. *Trends in Cognitive Sciences*, 17: 547-549.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, 1: 333-342.

Hilbig, B. E. (2010). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review*, 17: 923-930.

Hintze, A., Olson, R. S., Adami, C., and Hertwig, R. (2015). Risk sensitivity as an evolutionary adaptation. *Scientific reports*, 5: 8242.

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73: 538-40.

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, 84: 343-352.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science*, 290: 2261–2262.

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin,* 90: 197-217.

————————. (2005). The Challenge of representative design in psychology and economics. *Journal of Economic Methodology,* 12: 253-263.

Houston, A. (2012). Natural Selection and Rational Decisions. In S. Okasha and K. Binmore (Eds.), *Evolution and Rationality: Decisions, Co-operation and Strategic Behaviour*. Cambridge University Press.

Houston A. I., McNamara J. M., and Steer, M. D. (2007a) Violations of transitivity under fitness maximization. *Biology Letters*, 3: 365-367.

———————————————————. (2007b) Do we expect natural selection to produce rational behaviour? *Philosophical Transactions of the Royal Society of Science B*, 362: 1531-1543.

Hsee, C. K., and Hastie, R. (2006). Decision and experience: Why don't we choose what makes us happy? *Trends in Cognitive Sciences,* 10: 31-37.

Hsee, C. K., Hastie, R., and Chen, J. (2008). Hedonomics: Bridging decision research with happiness research. *Perspectives on Psychological Science*, 3: 224-243.

Hubbell, A. P., Mitchell, M. M., and Gee, J. C. (2001). The relative effects of timing of suspicion and outcome involvement on biased message processing. *Communication Monographs*, 68: 115-132.

Hurley, S. (2005). Social heuristics that make us smarter. *Philosophical Psychology*, 18: 585-612.

Hurley, S., and Nudds, M. (2006). *Rational animals?* Oxford University Press.

Hutchinson, J., and Gigerenzer, G. (2005). Simple heuristics and rules of thumb: Where psychologists and behavioral biologists might meet. *Behavioral processes*, 69: 97-124.

Jarvstad, A., Hahn, U., Warren, P. A., and Rushton, S. K. (2014). Are perceptuo-motor decisions really more optimal than cognitive decisions? *Cognition*, 130: 397-416.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Johnson-Laird, P. N. (1983). *Mental models*. Harvard University Press.

Johnson, D., and Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, 477: 317-320.

Johnson, D., and Levin, S. (2009). The tragedy of cognition: psychological biases and environmental inaction. *Current Science*, 97: 1593-1603.

Jolls, C., Sunstein, C. R., and Thaler, R. (1998). A behavioral approach to law and economics. *Stanford Law Review*, 50: 1471-1550.

Jones, M., and Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition. *Behavioural and Brain Sciences*, 34: 169-88.

Jones, S. K., Jones, T., and Frisch, D. (1995). Biases of probability assessment: A comparison of frequency and single-case judgments. *Organizational Behaviour and Human Decision Processes,* 61: 109-122.

Jung, R. E., and Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: converging neuroimaging evidence. *Behavioural and Brain Sciences*, 30: 135-154.

Kacelnik A. (2006). Meanings of rationality. In M. Nudds and S. Hurley (Eds.), *Rational animals?* Oxford University Press.

Kacelnik, A., Schuck-Pain, C., and Pompilio, L. (2006). Inconsistency in animal and human choice. In C. Engel and L. Daston (Eds.), *Is there value in inconsistency?* Nomos.

Kagel, C. J. (1987). Economics according to the rats (and pigeons too): What have we learned and what we hope to learn. In A. Roth (Ed.), *Laboratory experimentation in economics: Six points of view*. Cambridge University Press.

Kahneman, D. (1981). Who shall be the arbiter of our intuitions? *Behavioural and Brain Sciences*, 4: 339-340.

———————. (1997). New Challenges to the Rationality Assumption. *Legal Theory*, 3: 105-124.

Kahneman, D., and Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, and D. Kahneman, (Eds.), *Heuristics and biases: The psychology of intuitive thought*. Cambridge University Press.

———————————. (2005). A model of heuristic judgment. In K. J. Holyoak and R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press.

Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgment of representa- tiveness. *Cognitive Psychology*, 72: 430-454.

———————————. (1973). On the psychology of prediction. *Psychological Review,* 80: 237-251.

———————————. (1979). Prospect theory: An analysis of decision under risk. *Econometrica,* 47: 263-92.

———————————. (1982a). The psychology of preferences. *Scientific American,* 246: 160-173.

———————————. (1982b). On the Study of Statistical Intuitions. In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

———————————. (1984). Choices, Values and Frames. *American Psychologist*, 39, 344-350.

———————————. (1996). On the reality of cognitive illusions. *Psychological Review,* 103: 592–591.

Kassin, S. M. (2005). On the psychology of confessions: does innocence put innocents at risk? *American Psychologist*, 60: 215-228.

Katsikopoulos, K. V. (2009). Coherence and correspondence in engineering design: informing the conversation and connecting with judgment and decision-making research. *Judgment and Decision Making*, 4: 147-153.

Keeney, R., and Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press.

Kelly, T. (2003). Epistemic rationality as instrumental rationality: A critique. *Philosophy and Phenomenological Research*, 66: 612-640.

Kelly, D. and Roeddert, R. (2008) Racial cognition and the ethics of implicit bias. *Philosophy Compass*, 3: 522–540.

Kelman, M. (2013). Moral realism and the heuristics debate. *Journal of Legal Analysis*, 5: 339-397.

Kenrick, D. T., and Griskevicius, V. (2013). *The rational animal: How evolution made us smarter than we think*. Basic Books.

Kenrick, D. T., Griskevicius, V., Sundie, J. M., and Neuberg, S. L. (2009). Deep rationality: the evolutionary economics of decision-making. *Social cognition*, 27: 764-785.

Kenrick, D. T., and Keefe, R. C. (1992) Age preferences in mates reflect sex differences in reproductive strategies. *Behavioural and Brain Sciences*, 15: 75–91.

Kenrick, D. T., Li, Y. J., White, A. E., and Neuberg, S. L. (2012). Economic subselves: Fundamental motives and deep rationality. *Social Thinking and Interpersonal Behavior*, 14: 23-43.

Kenrick, D. T., Neuberg, S. L., Griskevicius, V., Becker, D. V., and Schaller, M. (2010). Goal-Driven Cognition and Functional Behavior The Fundamental-Motives Framework. *Current Directions in Psychological Science*, 19: 63-67.

Koehler, J. (1996). On conveying the probative value of DNA evidence: frequencies, likelihood ratios, and error rates. *University of Colorado Law Review,* 67: 859–886.

Koenigs, M., and Tranel, D. (2007). Prefrontal cortex damage abolishes brand-cued changes in cola preference. *Social cognitive and affective neuroscience*, 3: 1-6.

Kohler, W. (1927). *The mentality of apes*. London: Routledge & Kegan Paul.

Korsgaard, C. (1997). The Normativity of Instrumental Reason. In G. Cullity and B. Gaut (eds.), *Ethics and Practical Reason*. Clarendon Press.

Koscik,T. R., and Tranel, D. (2013). Abnormal causal attribution leads to advantageous economic decision-making: a neuropsychological approach. *Journal Cognitive Neuroscience*, 25: 1372–1382.

Kraut, R. E. (1980). Humans as lie detectors: Some second thoughts. *Journal of Communication*, 30: 209–216.

Kruglanski, A. W. (1996). Goals as knowledge structures. In P. M. Gollwitzer and J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behaviour*. Guilford Press.

———————. (2013). Only one? The default interventionist perspective as a unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8: 242-247.

Kruglanski, A. W., and Ajzen, I. (1983). Bias and error in human judgment. *European Journal of Social Psychology*, 13: 1–44.

Kruglanski, A. W., and Köpetz, C. (2009). What is so special (and nonspecial) about goals. In G. B. Moskowitz and H. Grant (eds.), *The psychology of goals*. Guilford Press.

Kruglanski, A. W., Pierro, A., Mannetti, L., Erb, H. P., and Chun, W. Y. (2007). On the parameters of human judgment. *Advances in Experimental Social Psychology*, 39: 255-303.

Kuncel, N. R., and Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*, 19: 339-345.

Langer, E. (1975) The Illusion of Control. *Journal of Personality and Social Psychology*, 32: 311-328.

Larrick, R., Nisbett, R., and Morgan, J. (1993). Who uses the cost–benefit rules of choice? Implications for the normative status of microeconomic theory. *Organizational Behaviour and Human Decision Processes*, 56: 331–347.

Lee, J. C. (2008). Epistemology by applied cognitive psychology and the "strong replacement" of normative psychology. *Philosophy of the Social Sciences*, 38: 55-75.

Lee, M. D., and Zhang, S. (2012). Evaluating the coherence of Take-the-best in structured environments. *Judgment and Decision Making*, 7: 360-372.

Lenton, A. P., Penke, L., Todd, P. M., and Fasolo, B. (2013). The heart has its reasons: Social rationality in mate choice. In R. Hertwig, U. Hoffrage, and the ABC Research Group (Eds.), *Simple heuristics in a social world*. Oxford University Press.

Lerner, J., and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125: 255-275.

Levine, T. R., Park, H. S., and McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communications Monographs*, 66: 125-144.

Levitt, S. D., and List, J. A. (2009). Field experiments in economics: The past, the present and the future. *European Economic Review*, 53: 1-18.

Lewis, D. (1981). Why ain'cha rich? *Nous*, 15: 377-380.

Lieder, F., Griffiths, T. L., and Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), *Advances in neural information processing systems*. MIT Press.

Lin, H. (2014). On the regress problem of deciding how to decide. *Synthese*, 191: 661-670.

List, J. A. (2008). Homo experimentalis evolves. *Science,* 321: 207-208.

Lopes, L. L. (1991). The rhetoric of irrationality. *Theory & Psychology*, 1: 65-82.

Lubinski, D. (2009). Cognitive epidemiology: With emphasis on untangling cognitive ability and socioeconomic status. *Intelligence*, 37: 625-633.

Lurz, R. W. (2011). *Mindreading animals: the debate over what animals know about other minds*. MIT Press.

Machery, E., and Cohen, K. (2012). An evidence-based study of the evolutionary behavioral sciences. *The British Journal for the Philosophy of Science*, 63: 177–226.

Maciejovsky, B., and Budescu, D. V. (2007). Collective induction without cooperation? Learning and knowledge transfer in cooperative groups and competitive auctions. *Journal of Personality and Social Psychology*, 92: 854-870.

Mackintosh, N. (2011). *IQ and human intelligence*. Oxford University Press.

Mandel, D. R. (2005). Are risk assessments of a terrorist attack coherent? *Journal of Experimental Psychology: Applied*, 11: 277-288.

———————. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General*, 143: 1185-1198.

Marewski, J. N., Gaissmaier, W., and Gigerenzer, G. (2010). We favour formal models of heuristics rather than lists of loose dichotomies: a reply to Evans and Over. *Cognitive Processing*, 11: 177-179.

Smith, J. M., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., and Wolpert, L. (1985). Developmental constraints and evolution: a perspective from the Mountain Lake conference on development and evolution. *Quarterly Review of Biology*, 60: 265-287.

McClintock, C. G., and Liebrand, W. B. (1988). Role of interdependence structure, individual value orientation, and another's strategy in social decision making: A transformational analysis. *Journal of personality and social psychology*, 55: 396-409.

McKay R., and Dennett, D. (2009) The evolution of misbelief. *Behavioural and Brain Sciences*, 32: 493-561.

McKenna, F. P., Stanier R. A. and Lewis, C. (1991) Factors underlying illusory self-assessment of driving skill in males and females. *Accident Analysis and Prevention*, 23: 45-52.

McKenzie, C. R. (2004). Framing effects in inference tasks—and why they are normatively defensible. *Memory & Cognition*, 32: 874-885.

McNamara, J. M., Stephens, P. A., Dall, S. R., and Houston, A. I. (2009). Evolution of trust and trustworthiness: social awareness favors personality differences. *Proceedings of the Royal Society B: Biological Sciences*, 276: 605-613.

McNeil, B. J., Pauker, S. G., Sox Jr, H. C., and Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *The New England Journal of Medicine*, 306: 1259-1262.

Mellers, B., Hertwig, R., and Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12: 269-275.

Mercier, H., and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, 34: 57-74.

Milkman, K. L., Chugh, D., and Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science*, 4: 379-383.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63: 81-97.

Minelli, A. (2009). *Forms of becoming: The evolutionary biology of development*. Princeton University Press.

————. (2010). Evolutionary developmental biology does not offer a significant challenge to the neo-Darwinian paradigm. In F. Ayala and R. Arp (Eds.), *Contemporary debates in the philosophy of biology*. Wiley- Blackwell.

Misyak, J. B., and Chater, N. (2014). Virtual bargaining: a theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369: 1-9.

Mohlin, E. (2012). Evolution of theories of mind. *Games and Economic Behavior*, 75: 299-318.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38: 379-387.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115: 502-517.

Moro, R. (2009). On the nature of the conjunction fallacy. *Synthese*, 171: 1-24.

Morton, A. (2010). Human bounds: rationality for our species. *Synthese*, 176: 5-21.

————. (2012). *Bounded thinking: Intellectual virtues for limited agents*. Oxford University Press.

Newell, B. R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences*, 9: 11-15.

Newell, B., Lagnado, D., and Shanks, D. R. (2007) *Straight choices: The psychology of decision making*. Psychology Press.

Newell, B. R., and Shanks, D. R. (2014). Unconscious influences on decision-making: A critical review. *Behavioral and Brain Sciences*, 37: 1-19.

Nichols, S., and Stich, S. P. (2003). *Mindreading: An integrated account of pretense, self-awareness, and understanding other minds*. Oxford University Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2: 175-220.

Nilsson, H. and P. Andersson. (2010). Making the seemingly impossible appear possible: Effects of conjunction fallacies in evaluations of bets on football games. *Journal of Economic Psychology*, 3: 172–180.

Nisbett, R. E., and Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, 32: 932-943.

Nisbett, R. E., and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs.

Nozick, R. (1963), *The Normative Theory of Individual Choice*. Garland Publishing.

————. (1969). Newcomb's problem and two principles of choice. In Rescher, N. (Ed.), *Essays in honor of Carl G. Hempel*. Springer.

————. (1993). *The Nature of Rationality*. Princeton University Press.

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review,* 101: 608-631.

————————————. (1996). Rational explanation of the selection task. *Psychological Review,* 103: 381-391.

————————————. (2007). *Bayesian rationality*. Oxford University Press.

Olsson, E. (2005). *Against Coherence: Truth, Probability, and Justification*. Oxford University Press.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11: 988-1010.

————————. (2013). A Case Study Dual-Process Theories of Higher Cognition—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8: 248-252.

————————. (2014). *Future-minded: The psychology of Agency and Control*. Palgrave Macmillan.

Over, D. E. (2000). Ecological Rationality and its Heuristics. *Thinking and Reasoning*, 6: 182–192.

————————. (2004). Rationality and the normative/descriptive distinction. In D. J. Koehler and N. Harvey (Eds.), *Blackwell handbook of judgment and decision making*. Blackwell Publishing.

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, 81: 181–192.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Penke, L. (2010). Bridging the gap between modern evolutionary psychology and the study of individual differences. In D. M. Buss and P. H. Hawley (Eds.), *The evolution of personality and individual differences*. Oxford University Press.

Perilloux, C. (2014). (Mis)reading the Signs: Men's Perception of Women's Sexual Interest. In Weekes-Shackelford, V. A., and Shackelford, T. K. (Eds.), *Evolutionary Perspectives on Human Sexual Psychology and Behavior*. Springer.

Perilloux, C., Easton, J. A., and Buss, D. M. (2012). The misperception of sexual interest. *Psychological Science*, 23: 146-151.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, 17: 407–413.

Peterson, C. R., and Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68: 29–46.

Peterson, C., and Uleha, Z. (1964). Uncertainty, inference difficulty and probability learning. *Journal of Experimental Psychology*, 67: 523–530.

Pettit, P., and Sugden, R. (1989). The backward induction paradox. *The Journal of Philosophy*, 86: 169-182.

Phillips, L. D., and Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72: 346-354.

Piattelli-Palmarini, M. (1991). Probability blindness: Neither rational nor capricious. *Bostonia*, 28-35.

Plott, C. (1996). Rational Individual Behavior in Markets and Social Choice Processes: the Discovered Preference Hypothesis. In K. J. Arrow, M. Perlman and C. Schmidt (Eds.), *The Rational Foundations of Economic*. Basingstoke.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, 17: 407-413.

Pigliucci, M. (2010). A misguided attack on evolution. *Nature*, 464: 353-354.

Pigliucci, M., and Müller, G. B. (2010). *Evolution, the extended synthesis*. MIT Press.

Politzer, G. (2004). Reasoning, judgement and pragmatics. In N. Noveck and D. Sperber (Eds.), *Experimental pragmatics*. Palgrave.

Politzer, G., and Noveck, I. A. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research*, 20: 83-103.

Pothos, E. M., and Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, 36: 255-274.

———————————. (2014). In search for a standard of rationality. *Frontiers in psychology*, 5.

Prinz, J. (2008). Empirical Philosophy and Experimental Philosophy. In J. Knobe and S. Nichols (eds.), *Experimental Philosophy*. Oxford University Press.

Prinz, J. J., and Barsalou, L. W. (2000). Steering a course for embodied representation. In E. Dietrich and A. B. Markmam (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines*. Lawrence Erlbaum Associates.

Powell, R. (2012). The future of human evolution. *British Journal for the Philosophy of Science,* 63: 145–175.

Raab, M., and Gigerenzer, G. (2005). Intelligence as smart heuristics. In R. J. Sternberg, J. Davidson, and J. Pretz (Eds.), *Cognition and intelligence*. Cambridge University Press.

Read, D. (2005). Monetary Incentives, What Are They Good For? *Journal of Economic Methodology*, 12: 265–76.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, 253: 980–986.

Reimer, T., and Katsikopoulos, K. (2004). The use of recognition in group decision-making. *Cognitive Science* 28: 1009–29.

Reyna, V. F., and Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, 18: 89-107.

———————————. (2007). The importance of mathematics in health and human judgment: Numeracy, risk communication, and medical decision-making. *Learning and Individual Differences*, 17: 147-159.

Reyna, V. F., and Lloyd, F. (1997). Theories of false memory in children and adults. *Learning and Individual Differences*, 9: 95-123.

Reyna, V. F., Nelson, W. L., Han, P. K., and Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision-making. *Psychological Bulletin*, 135: 943-973.

Rich, P. (2014). Comparing the axiomatic and ecological approaches to rationality: fundamental agreement theorems in SCOP. *Synthese*, 1-19.

Richardson, H. S., (1994). *Practical Reasoning about Final Ends*. Cambridge University Press.

Richter, T., and Spath, P. (2006). Recognition is used as one cue among others in judgment and decision-making. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 32: 150–162.

Rieskamp, J., and Reimer, T. (2007). Ecological rationality. In R. Baumeister, and K. Vohs (Eds.), *Encyclopedia of social psychology*. SAGE Publications.

Rini, R. A. (2015). Psychology and the aims of normative Ethics. *Handbook of Neuroethics*, 149-168.

Rips, L. J. (2002). Circular reasoning. *Cognitive Psychology*, 26: 767-795.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13: 279-301.

Rutter, D. R., Quine, L., and Albery, I. P. (1998) Perceptions of risk in motorcyclists: Unrealistic optimism, relative optimism, and predictions of behaviour. *British Journal of Psychology*, 89: 681-697.

Rysiew, P. (2008). Rationality disputes–psychology and epistemology. *Philosophy Compass*, 3: 1153-1176.

Sage, W. (2004) Truth-reliability and the evolution of human cognitive faculties. *Philosophical Studies*, 117: 95-106.

Samuels, R., and Stich, S. (2004). Rationality and psychology. In A. Mele and P. Rawling (Eds.), *The Oxford handbook of rationality*. Oxford University Press.

Samuels, R., Stich, S, and Bishop, M. (2002). Ending the rationality wars: how to make disputes about human rationality disappear? In R. Renee (Ed.), *Common Sense, Reasoning and Rationality*. Oxford University Press.

Samuels, R., Stich, S., and Faucher, L. (2004). Reasoning and Rationality. In I. Niiniluoto, M. Sintonen, and J. Wolenski (Eds.), *Handbook of Epistemology*. Kluwer.

Scheibehenne, B., Wilke, A., and Todd, P. M. (2011). Expectations of clumpy resources influence predictions of sequential events. *Evolution and Human Behavior*, 32: 326-333.

Schuck-Paim, C., and Kacelnik, A. (2002). Rationality in risk-sensitive foraging choices by starlings. *Animal Behaviour*, 64: 869–879.

Schick, F. (1991). *Understanding Action*. Cambridge University Press.

Schulz, A. W. (2008). Risky Business: Evolutionary Theory and Human Attitudes toward Risk—A Reply to Okasha. *The Journal of Philosophy*, 104: 156-165.

————————. (2011a) Gigerenzer's evolutionary arguments against rational choice theory: an assessment. *Philosophy of Science*, 78: 1272-1282.

————————. (2011b). Simulation, simplicity, and selection: an evolutionary perspective on high-level mindreading. *Philosophical Studies*, 152: 271-285.

———————. (forthcoming). Preferences vs. Desires: Debating the Fundamental Structure of Conative States." *Economics and Philosophy*.

Schurz, G. (2014). Cognitive Success: Instrumental Justifications of Normative Systems of Reasoning. *Frontiers in Psychology*, 5: 625.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology*, 26: 123-162.

Searle, J. R. (2001). *The rationality of action*. MIT Press.

Sedikides, C., Ariely, D., and Olsen, N. (1999). Contextual and Procedural Determinants of Partner Selection: Of Asymmetric Dominance and Prominence. *Social Cognition*, 17: 118-139.

Sen, A. (1993). Internal consistency of choice. *Econometrica,* 61: 495–521.

———. (2002). *Rationality and freedom*. Harvard University Press.

Shafir, E., and LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, 53: 491-517.

Shapiro, L. (2010). James Bond and the Barking Dog: Evolution and Extended Cognition. *Philosophy of Science*, 77: 400-418.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics,* 69: 99–118.

———————. (1957). *Models of man: Social and rational*. New York: Wiley.

———————. (1973). The structure of ill-structured problems. *Artificial Intelligence,* 4: 181–200.

———————. (1979). Rational decision making in business organizations. *The American Economic Review*, 69: 493-513.

———————. (1983). *Reason in human affairs*. Stanford University Press.

———————. (1990). Alternative visions of rationality. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches*. Cambridge University Press.

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research* 16: 158-174.

Simpson, G. (1963). *This view of life: the world of an evolutionist*. Harper.

Sinnott-Armstrong, W., Young, L., and Cushman, F. (2010). Moral intuitions. In J. M. Doris and The Moral Psychology Research Group (Eds.), *The moral psychology handbook*. Oxford University Press.

Slezak, P. (1999). Situated cognition: empirical issue, paradigm shift or conceptual confusion? In J. Wiles and T. Dartnal (Eds.) *Perspectives on cognitive science*, Vol. 2. Ablex.

Sloman, S. A., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behaviour and Human Decision Processes*, 29: 296-309.

Slovic, P., and Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behaviour and Human Performance*, 6: 649–744.

Smith, E. A., Borgerhoff Mulder, M., and Hill, J. (2001). Controversies in the evolutionary social sciences: A guide for the perplexed. *Trends in Ecology and Evolution*, 16: 128–135.

Smith, V. (2003). Constructivist and Ecological Rationality in Economics. *American Economic Review*, 93: 465-508.

Sober, E. (2010). Natural Selection, Causality, and Laws: What Fodor and Piattelli-Palmarini Got Wrong. *Philosophy of Science*, 77: 594-607.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Psychology Press.

———————. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. University of Chicago Press.

———————. (2011a). Normative models in psychology are here to stay. *Behavioral and Brain Sciences*, 34: 268-269.

———————. (2011b). *Rationality and the reflective mind*. Oxford University Press.

———————. (2013). Why humans are (sometimes) less rational than other animals: Cognitive complexity and the axioms of rational choice. *Thinking & Reasoning*, 19: 1-26.

Stanovich, K. E., and West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, 127: 161-188.

———————————. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23: 645-665.

———————————. (2003) Evolutionary versus instrumental goals: how evolutionary psychology misconceives human rationality. In: Over E. (ed.) *Evolution and the psychology of thinking: the debate*. Psychology Press.

———————————. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94: 672-695.

———————————. (2014). What intelligence tests miss. *The Psychologist*, 27: 80-83.

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

Stein, E. (1996). *Without good reason: The rationality debate in philosophy and cognitive science*. Clarendon Press.

Stephens, C. L. (2001). When is it selectively advantageous to have true beliefs? Sandwiching the better safe than sorry argument. *Philosophical Studies*, 105: 161-189.

Stephens, D. W., and Krebs, J. R. (1986). *Foraging theory*. Princeton University Press.

Sterelny, K. (2000). Development, evolution, and adaptation. *Philosophy of Science*, 67: S369-S387.

————————. (2003). *Thought in a hostile world: The evolution of human cognition*. Oxford: Blackwell.

Stern, R. (2004). Does 'ought' imply 'can'? And did Kant think it does? *Utilitas*, 16: 42-61.

Sternberg, R. J. (1997). *Successful intelligence*. Plume.

Stevens, J. R. (2008). The evolutionary biology of decision making. In: Engel C, Singer W (Eds.), *Better than conscious?* MIT Press.

————————. (2010) Rational decision making in primates: the bounded and the ecological. In: Platt M. L., Ghazanfar A. A. (ed.) *Primate neuroethology*. Oxford University Press.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, 35: 401-426.

Stich, S. (1990) *The fragmentation of reason*. Cambridge University Press.

Stolarz-Fantino, S., Fantino, E., Zizzo, D. J., and Wen, J. 2003. The conjunction fallacy: new evidence for robustness. *American Journal of Psychology*, 116: 15–34.

Sturm, T. (2012). The "rationality wars" in psychology: Where they are and where they could go. *Inquiry*, 55: 66-81.

Sugden, R. (2006). Hume's non-instrumental and non-propositional decision theory. *Economics and Philosophy*, 22: 365-391.

Sunstein, C. R. (1997). Behavioral analysis of law. *University of Chicago Law Review,* 64: 1175-1195.

Sutton, J., Harris, C., Keil, P., and Barnier, A. (2010). The psychology of memory, extended cognition and socially distributed remembering. *Phenomenology and the Cognitive Science*, 9: 521–560.

Svenson, O. (1981) Are we all less risky and more skilful than our fellow drivers? *Acta Psychologica*, 47: 143–148.

Symons, D. (1992). On the Use and Misuse of Darwinism in the Study of Human Behavior. In J. H. Barkow, L. Cosmides, and J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind*. Basic Books.

Taylor, S. E., and Brown, J. D. (1988) Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103: 193–210.

Taylor S. E., Lerner J. S., Sherman, D. K., Sage R. M., and McDowell, N. K. (2003) Are self-enhancing cognitions associated with healthy or unhealthy biological profiles? *Journal of Personality and Social Psychology*, 85: 605-615.

Téglás, E., Girotto, V., Gonzalez, M., and Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences*, 104: 19156–19159.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., and Bonatti, L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 332: 1054–1059.

Tentori, K., Bonini, N., and Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science,* 28: 467–477.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1: 39-60.

Thaler, R. H., and Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thompson, S. C, Armstrong, W. and Thomas, C. (1998) Illusions of control, underestimations, and accuracy: A control heuristic explanation. *Psychological Bulletin*, 123: 143-161.

Todd, P. (2001). Fast and frugal heuristics for environmentally bounded minds. In: Gigerenzer G., and Selten R. (Eds.) *Bounded rationality: the adaptive toolbox*. MIT Press.

Todd, P. M., and Gigerenzer, G. (2000) Précis of Simple Heuristics That Make Us Smart. *Behavioural and Brain Sciences*, 23: 727-741.

———————————. (2012). *Ecological rationality: Intelligence in the world*. Oxford University Press.

Todd, P. M., and Miller, G. F. (1999). From pride and prejudice to persuasion: Satisficing in mate search. In G. Gigerenzer, P. M. Todd, and the ABC Research Group (eds.), *Simple Heuristics That Make Us Smart*. Oxford University Press.

Tooby, J., and Cosmides, L. (1990). On the universality of human nature and the uniqueness of the individual: The role of genetics and adaptation. *Journal of Personality,* 58: 17-67.

Toplak, M. E., West, R. F., and Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39: 1275-1289.

Trivers, R. L. (1972) Parental investment and sexual selection. In Campbell B. (ed.), *Sexual selection and the descent of man: 1871–1971*. Aldine, Chicago.

————————. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books.

Trommershäuser, J., Landy, M. S., and Maloney, L. T. (2006). Humans rapidly estimate expected gain in movement planning. *Psychological Science*, 7: 981–988.

Trommershäuser, J., Maloney, L. T., and Landy, M. S. (2008). Decision-making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, 12: 291–297.

Tversky, A., and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211: 453-438.

———————————————. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review,* 90: 293–315.

———————————————. (1986). Rational choice and the framing of decisions. *The Journal of Business,* 59: S251–78.

Ulehla, Z. J. (1966). Optimality of perceptual decision criteria. *Journal of Experimental Psychology*, 71: 564–569.

Vallinder, A., and Olsson, E. J. (2014). Trust and the value of overconfidence: a Bayesian perspective on social network communication. *Synthese*, 191: 1991-2007.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, 92: 345-381.

Van Osselaer, S. M., Ramanathan, S., Campbell, M. C., Cohen, J. B., Dale, J. K., Herr, P. M., and Tavassoli, N. T. (2005). Choice based on goals. *Marketing Letters*, 16: 335-346.

Varela, F., Thompson, E., and Rosch, E. (1991). *The Embodied Mind*. MIT Press,

Vernon, P. E. (1964). *The structure of human abilities*. Methuen.

Vranas, P. (2000). Gigerenzer's Normative Critique of Kahneman and Tversky. *Cognition*, 76: 179-193.

————————. (2007). I Ought, Therefore I Can. *Philosophical Studies*, 136: 167 - 216.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and implications for professional practice*. John Wiley & Sons.

Wagner, A. (2005). Robustness and evolvability in living systems. *Princeton studies in complexity*.

Wallin, A. (2013) A peace treaty for the Rationality Wars? External validity and its relation to normative and descriptive theories of rationality. *Theory & Psychology*, 23: 458-478.

Wason, P. (1966). Reasoning. In Foss, B. M. (Ed.), *New horizons in psychology*. Penguin.

Watts, D. (2003). *Six degrees*. Norton.

Wedell, D. H., and Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: effects of response mode, conceptual focus and problem type. *Cognition,* 107: 105-136.

Weinberg, J. M., Nichols, S., and Stich, S. (2001). Normativity and Epistemic Intuitions. *Philosophical Topics*, 29: 429-460.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., and Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human behavior*, 22: 603-647.

Wendt, A. (2015). *Quantum Mind and Social Science Unifying Physical and Social Ontology*. Cambridge University Press.

West, R. F., Toplak, M. E., and Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, 100: 930-941.

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, 25: 41-78.

Wilcox, N. (1993) Lottery choice, incentives, complexity and decision time. *Economic Journal*, 103: 1397–417.

Wilke, A., and Mata, R. (2012) Cognitive bias. In V. S. Ramachandran (Ed.), *Encyclopaedia of Human Behaviour* (2nd Edition). Elsevier.

Wilke, A., and Todd, P. M. (2012). The evolved foundations of decision-making. In M. K. Dhami, A. Schlottmann, and M. Waldmann (Eds.), *Origins of judgment and decision making*. Cambridge University Press.

Wilson, D. S. (2010). *Darwin's cathedral: Evolution, religion, and the nature of society*. University of Chicago Press.

Wilson, R. (2010). Meaning making and the mind of the externalist. In R. Menary (Ed.), *The extended mind*. MIT Press.

Wilson, T. D., and Brekke, N. C. (1994) Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, 116: 117-142.

Winograd, T., and Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Intellect Books.