CHAPTER TWO

What Does It Mean to be Biased: Motivated Reasoning and Rationality

Ulrike Hahn[*,1], **Adam J.L. Harris**[†]
[*]Department of Psychological Sciences, Birkbeck, University of London, London, United Kingdom
[†]Department of Cognitive, Perceptual & Brain Sciences, University College London, London, United Kingdom
[1]Corresponding author: e-mail address: u.hahn@bbk.ac.uk

Contents

Abstract

In this chapter, we provide a historical overview of research on bias in human cognition, ranging from early work in psychology through the detailed, quantitative examinations of belief revision in the 1960s, the Heuristic and Biases program initiated by Kahneman and Tversky, and bias focused research in personality and social psychology. Different notions of "bias" are identified and compared with the notion of bias in statistics, machine learning, and signal detection theory. Comparison with normative models then forms the basis for a critical look at the evidence that people succumb to motivated reasoning aimed at enabling them "to believe what they want to believe."

Psychology of Learning and Motivation, Volume 61
ISSN 0079-7421
http://dx.doi.org/10.1016/B978-0-12-800283-4.00002-2

1. THE NOTION OF BIAS

A reader venturing into the psychological literature about human biases soon realizes that the word "bias" means many things to many people. This holds not just for its wider connotations, but even its immediate meaning. Consequently, it seems necessary to start with a survey of the term's usage.

In everyday use, the term "bias" refers to a lack of impartiality or an undue preference: bias is "an inclination or prejudice for or against one person or group, especially in a way considered to be unfair" (Oxford English Dictionary), or "a tendency to believe that some people, ideas, etc., are better than others that usually results in treating some people unfairly" (Merriam Webster). However, even dictionary definitions contain related meanings that lack the negative connotation, with "bias" being described also as "a strong interest in something or ability to do something" (Merriam Webster).

Already apparent in these definitions of everyday use are a number of fundamental distinctions: whether bias is a property of beliefs or of decisions, and whether or not it is inherently negative or "wrong."

However, these are not the only dimensions of variation that may structure the debate within the psychological literature. The word "bias" also has sharpened, more technical, meanings—in particular in statistics—that are also sometimes intended in research on bias. Statistical treatments also provide very balanced consideration of when being "biased" might be good, so we will introduce these technical meanings in more detail. We start, however, by an overview of the concept of "bias" within psychological research.

1.1. "Bias" in Psychology

Without a more thorough historical analysis than we are willing (or able) to conduct, any overview of the enormous wealth of research on biases and its reception within cognitive and social psychology will necessarily remain subjective and incomplete. Our main goal is to identify broad contrasts and key dimensions of variation thus setting the stage for a more detailed look at a particular class of biases—indicative of "motivated reasoning"—in the second half of this chapter.

1.1.1 Origins

Interest in biases within human cognition developed early in psychology. Vaughan's (1936) book "General Psychology" contains a chapter titled "The Importance of Bias." On the definition of bias, Vaughan writes:

> *A bias is a slant or bent, a pointing of the mental life toward certain views and reactions. Consciously, a bias is a point of view; behavioristically, it is a posture, a set, a preparedness for acting, thinking, or judging, in a definite manner. A bias is an attitude—an anticipation—a prejudice which may manifest itself in overt behavior or in thoughts and feelings about behavior. Very often the determining tendency operates unconsciously, that is, without the individual's being aware of the motive fundamentally responsible for his thinking or action. (p. 211)*

Vaughan speaks of biases in perception, memory, judgment, belief, and choice. Some of his evidence is informal and anecdotal; however, psychological research at that time already possessed empirical demonstrations of judgment biases (e.g., Macdougall, 1906—savor the hand drawn graphs!), attitudinal biases and their impact on memory (e.g., Levine & Murphy, 1943), attentional biases in perception (e.g., Wolff, 1933), response biases (see e.g., Landahl, 1939), and a sizeable literature on perceptual illusions (see e.g., Pierce, 1901).

One also already finds in this literature both an interest in studying biases with a view to allowing human beings to overcome them and with a view to studying biases as a means of coming to understand underlying mechanisms—two motivations that are reiterated in the literature on biases time and again (e.g., Gregg & Sedikides, 2004; Kahneman, 2000; Kahneman & Tversky, 1996; Kunda, 1990). Finally, the early literature already distinguishes between a notion of bias as filtering or selectivity in a very general sense and a more specific sense of bias as a distortion. Macdougall (1906) writes:

> *. . .selective attention working under the guidance of our organic interests operates upon the materials and processes of the external world, adding accentuation and emphasis, seizing upon and preserving certain elements which we call pleasing or important, and relegating the rest to obscurity or oblivion. Often the account in which this recasting results is unrecognizable by a fellow-observer of the event. The existence of subjective bias is thus not an incidental error in our observations but is fundamental to the very character of the human mind. We can conceive its elimination only in an absolutely dispassionate consciousness devoid of feeling and purpose.*
>
> *This universal bias roots in the fact that at each moment of our experience some one interest is for the time being paramount, and determines both the objects which shall be attended to and the interpretation which they shall receive. (p. 99)*

MacDougall considers such subjective selection and emphasis "to pervade all mental activities, perceptive, imaginative, and rationalizing," but also, once acknowledged in its existence, not to be a concern. Indeed, he considers it "the basis of intelligibility in the world and of a rational adjustment to its

changes" that "the apprehension of that world varies from moment to moment in dependence upon transitions in the point of view and present purpose of the beholder" (p. 100).

An altogether different matter, however, is what he calls "bias of the second order," namely *distorted evaluations* of our necessarily selective perceptions of the world. These too he considers to be ubiquitous ("as pervasive as gravity," in fact), but of considerable practical consequence, because the distortions arise through "factors of evaluation of whose presence we are unaware at the moment of judgment" (p. 100).

He takes such evaluative distortions to arise at all levels of the system, from sensation through perception to memory and judgment, through to evaluation of complex conceptual objects. And in an assessment that could not have foreseen better the course of future research on this topic, he notes the ever-increasing difficulty of study as one moves through this list.

1.1.2 The 1960s: Wason's Confirmation Bias in Rule Induction

Given the range of different types of bias suggested by early psychological research, it may seem somewhat surprising that one of the most famous of all cognitive biases does not fit neatly into any of the categories mentioned so far. Peter Wason's (1960) paper "On the failure to eliminate hypotheses in a conceptual task" introduced the bias that has probably attracted the most enduring interest of all cognitive biases: the so-called confirmation bias. In his study, participants' task was to correctly infer a rule governing triplets of numbers (e.g., 2 4 6, and the underlying rule "increasing in magnitude") by generating query-triplets for which the experimenter indicated whether or not they conform to the rule. Wason's finding was that a proportion (though by no means all) of his participants sought to obtain evidence for what, if confirmed, would be positive instances, as opposed to negative instances. This tendency to "seek evidence that would confirm" violates the (then dominant) Popperian prescription of the need to seek falsification in the testing of scientific hypotheses (Popper, 1959), and was thus taken to fail the standard for "rational inference."

Given that it is about evidence selection, this "confirmation bias" seems closest to an attentional bias. It is not about suppression of particular content or information (e.g., attending to color as opposed to shape), but about strategy (i.e., deciding where to look): What kinds of questions should we ask of the world in order to determine the accuracy of our beliefs? The actual outcome of that query may turn out to confirm or to disconfirm our beliefs; hence, a "positive test strategy" must be distinguished from confirmation

or disconfirmation of the hypothesis itself. Nor need participants have any actual psychological desire to confirm the hypothesis for which they seek evidence (Wason, 1962; Wetherick, 1962), and the majority in Wason's study ultimately managed to infer the correct rule.

Nevertheless, "confirmation bias" has come to provide an umbrella term for a number of distinct ways in which beliefs and/or expectations influence both the selection, retention, and evaluation of evidence (see Nickerson, 1998, for a review). Nickerson (1998) lists under the banner of "confirmation" bias a wealth of distinct phenomena, drawn from both cognitive and social psychology, which we have collated here in Table 2.1.

In fact, these "subbiases" constitute the majority of phenomena listed by Baron (2008) under the header of "motivated bias," that is, biases reflecting "myside bias" or "wishful thinking." Confirmation bias has thus expanded from a particular type of search strategy to a concept considerably overlapping with the notion of "motivated reasoning," even though the original phenomenon contains no scrutiny of "motivation" whatsoever.

For the moment, it suffices to note that gathering evidence via search (like the retention of "evidence" in memory) necessarily has quite different standards for evaluation than does a distorting evaluation of evidence (or "secondary bias," in MacDougall's terms). In fact, such standards are not at all trivial to specify, and it is by no means enough to demonstrate merely that on occasion "something goes wrong."

1.1.3 The 1960s: Conservatism

Clear evaluative standards were, however, present in a wealth of research in the 1960s that examined carefully people's belief revision in light of new evidence. This line of research typically used "bookbags" and "pokerchips," that is, bags with varying compositions of colored chips (e.g., 60% red and 40% blue for one bag, and 40% red and 60% blue for the other). Participants then saw samples drawn from one of these bags and indicated their new, revised, degree of belief in the composition of that bag (e.g., that the bag was the one with predominantly blue chips). This paradigm allowed for careful quantitative evaluation of the extent to which participants' belief revision matched the prescriptions of Bayes' rule as a norm for updating beliefs (e.g., Peterson & Miller, 1965; Peterson & Uleha, 1964; Peterson, Schnieder, & Miller, 1965; Peterson, Uleha, Miller, & Bourne, 1965; Phillips & Edwards, 1966; see Peterson & Beach, 1967; Slovic & Lichtenstein, 1971, for reviews).

Table 2.1 Phenomena That Have Been Brought Under the Header of "Confirmation Bias"

1. *Hypothesis-determined information seeking and interpretation*	1.1 *Restriction of attention to a favored hypothesis.*	Considering only $P(D\|H)$ and not $p(D\|\text{not}H)$, for example, Doherty, Mynatt, Tweney, and Schiavo (1979)—sometimes referred to as *pseudodiagnosticity-bias*; but see Crupi, Tentori, and Lombardi (2009)
	1.2 *Preferential treatment of evidence supporting existing beliefs.*	*My-side bias*: tendency to produce reasons for favored side, for example, Baron (1995)
	1.3 *Looking only or primarily for positive cases.*	Tendency to ask questions for which answer would be "yes" if hypothesis were true: Wason (1960)
	1.4 *Overweighting positive confirmatory instances.*	For example, Gilovich (1983)
	1.5 *Seeing what one is looking for.*	For example, effects of expectations on social perception Kelley (1950); but Lenski and Leggett (1960) general tendency to respond to questions in acquiescence to interrogator hypothesis.
	1.6 *Remembering what one expects*	Eagly, Chen, Chaiken, and Shaw-Barnes (1999)
	1.7 *Illusory correlation*	Chapman and Chapman (1967), but see Fiedler and Krueger (2011)
2. Wason selection task and formal reasoning		Failure to pursue falsificationist strategy in context of conditional reasoning, Wason (1968); but see Oaksford and Chater (1994)
3. The primacy effect and belief persistence		*Resistance of* a belief or opinion to change once formed Pitz, Downing, and Reinhold's (1967) inertia effect; Lord, Ross, and Lepper (1979) "biased assimilation"
4. Overconfidence and the illusion of validity		For example, Lichtenstein and Fischhoff (1977); but see also Erev, Wallsten, and Budescu (1994)

Categories follow Nickerson's (1998) as do most of the examples, though newer references have been given in some cases.

The main finding of this research was that people responded in qualitatively appropriate ways to evidence, but—quantitatively—changed their beliefs less than the normative prescription of Bayes' rule mandated. In other words, their belief revision was what researchers called "*conservative*": people extracted less certainty from the evidence than it justified. Conservatism was found not just in relation to the diagnosticity of evidence but also to manipulations of the prior probability of the hypothesis in advance of the data, and affected not just belief revision but also the extent to which response criteria were shifted in normatively appropriate ways in signal detection tasks (which we discuss in more detail below; see e.g., Peterson & Beach, 1967; Ulehla, 1966 for further references).

These systematic deviations from optimal responding did not, however, lead researchers to form a negative conclusion of human rationality. In fact, the conclusion was that probability theory, which provides optimal models for making inferences under conditions of uncertainty, provides "a good first approximation for a psychological theory of inference" (Peterson & Beach, 1967, p. 42).

It is worth mentioning two sets of studies within the "bookbag and pokerchip" tradition that figure in discussions of motivated reasoning. First, after sequences of conflicting evidence that should have "cancelled out," participants' judgements did not necessarily return fully to the point of origin (see e.g., Pitz, 1969b; Peterson & DuCharme, 1967; Pitz et al., 1967; but see also Peterson et al., 1968), a phenomena dubbed the "inertia effect."

Second, under conditions where participants need to "purchase" information in order to reach a judgment, they purchased less information than they "should" (if they were maximizing expected value) and, hence, in a sense "jump to conclusions" (e.g., Green, Halbert, & Minas, 1964; Pitz, 1969a; though the reverse has also been found, see Tversky & Edwards, 1966; Wendt, 1969). This tendency to "under-sample" has been replicated many times since (see e.g., Fiedler & Kareev, 2006; Hertwig, Barron, Weber, & Erev, 2004). At first glance, it stands in seeming conflict with conservatism in belief revision, with people seeming simultaneously both too cautious and too decisive in their information evaluation. Such findings, in which people seem prone simultaneously to "opposing" biases have been a regular feature of the literature on biases ever since (see e.g., Table 1 in Krueger & Funder, 2004).

Though undersampling, like the inertia effect, has been assimilated into confirmation bias and motivated reasoning, cognitive psychologists have

recently argued that the tendency to select small samples reflects that sample proportions in small samples are exaggerated and may thus be easier to "read-off," an advantage that potentially comes at little cost (e.g., Fiedler & Kareev, 2006; Hertwig & Pleskac, 2008, 2010; on the wider role of small samples in judgment and decision see also, Hahn, 2014). Moreover, these benefits should be even more pronounced if evaluation of larger samples is conservative.

Despite a wealth of evidence, the causes of the pervasive conservatism observed in participants' judgments have never been fully resolved (see Erev et al., 1994). Edwards (1968) distinguished two possibilities: misaggregation and misperception. Participants could be misaggregating in their calculations of revised (posterior) degrees of belief; in keeping with this it was found that inference often seemed close to optimal with a single datum, deviating more strongly only as the amount of evidence increased (see e.g., DuCharme & Peterson, 1968; but see also DuCharme, 1970). Alternatively, participants could be misperceiving the diagnostic value of evidence. In keeping with this, participants (mis)perceived the sampling distributions from which their evidence was drawn to be flatter than they actually were (see e.g., Peterson, DuCharme, & Edwards, 1968; Wheeler & Beach, 1968). Training people to provide more veridical estimates of the underlying sampling distributions decreased conservatism. Furthermore, people's belief revision in general was typically found to be better predicted by their own subjective estimates of data characteristics, than by objective values; in other words, the internal consistency of people's probability judgments exceeded the correspondence of those judgments with objective, environmental values (see e.g., Peterson & Beach, 1967; Peterson, Schnieder, et al., 1965; Peterson, Uleha, et al., 1965, for further references); subjectively, people were "more Bayesian" than the degree of match between their judgments and the evidence suggested.

While there is thus both evidence in favor of misaggregation and in favor of misperception, neither factor explains all aspects of the data. Hence, other factors seem to play a role, including response bias. DuCharme (1970) found participants' responses to be optimal within a limited range either side of the initial, experimenter defined odds (i.e., the ratio between the probabilities of the two competing hypotheses under consideration).[1] Beyond this range, responses became increasingly conservative indicating a reluctance to move too far beyond whatever initial odds the experimenter provided (a reluctance which may reflect an everyday world in which very diagnostic evidence is rare). Within that range, however, responses showed neither misaggregation nor misperception.

[1] DuCharme (1970) found this range to correspond to log-odds of +/−1.

This fact argues against an explanation whereby conservatism is simply an artifactual result of a failure by participants to understand a complex experimental task (but see, e.g., Slovic & Lichtenstein, 1971). However, it has been demonstrated that the addition of random error to judgments may be one source of conservatism (e.g., Erev et al., 1994), and, in keeping with this, several of the studies that provided manipulations that reduced conservatism (e.g., Phillips & Edwards, 1966; Wheeler & Beach, 1968) reported reductions in variability.

In the 1970s, research in this area briefly turned to simple features of the evidence (such as sample proportion within the evidence seen) that participants might be tracking (see e.g., Manz, 1970; Slovic & Lichtenstein, 1971), before interest in the paradigm eventually waned.

This may to a good part be attributed to the fact that the optimistic assessment of human rationality soon gave way to a more dire assessment in the wake of Tversky and Kahneman's so-called Heuristics and Biases program.

Where authors such as Peterson and Beach (1967) were not only positive about the descriptive utility of probability and decision theory (and, in fact, anticipated the extension of their application to other aspects of human cognition), the project of "statistical man" or "man as an intuitive statistician" received a severe blow with Kahneman and Tversky's "Heuristics and Biases program" (e.g., Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1974).

1.1.4 Heuristics and Biases

This research tradition soon came to dominate cognitive psychological research on bias. It focused on (probability) judgment and decision-making, and "bias" in this context means systematic deviation from the (putative) normative standards of probability and expected utility theory. Systematic violations of these standards should give rise to outcomes that are inferior, either in terms of judgmental accuracy or goal attainment. The overarching interest in such systematic violations was motivated by a desire to find descriptively accurate characterizations of human judgment and decision-making that give insight into underlying mechanisms and processes. It has been much repeated within this tradition that biases may serve the same role of guiding the development of process theories as visual illusions had in the study of perception (see e.g., Kahneman & Tversky, 1982, 1996; Tversky & Kahneman, 1974).

Unlike in the study of perception, however, that promise has, to date, remained largely unfulfilled, and critics maintain that the paradigm has provided little more than a fairly haphazard list of supposed cognitive frailties

(see e.g., Krueger & Funder, 2004). Not only has the study of judgment and decision-making reached nowhere near the maturity of perception, but also both the empirical adequacy and explanatory power of the heuristics that supposedly underlie these biases have been severely doubted (see e.g., Gigerenzer, 1991, 2006).

Heuristics are procedures that are not guaranteed to succeed (see e.g., Newell, Shaw, & Simon, 1958) but that provide often highly effective shortcuts—in effect "rules of thumb." Consequently, heuristics, by definition, will occasionally—and systematically—be wrong. Heuristics thus bring with them "bias" (in the sense of systematic inaccuracy) by definition (see also, Kahneman, 2000).

More specifically, heuristics will be only partially correlated with true values. Where and when deviations occur depends on the nature of the heuristic. Substituting an easy to track property such as "availability" or "recognition" for the true determinants of some environmental frequency, for example, will overweight that property and neglect other, genuinely predictive, sources of information. This means the glass is half full: one may stress the fact of deviation; alternatively one may stress the fact that the heuristic often leads to the correct response given the actual environment in which the agent operates and that it does so in a computationally simple fashion, thus providing some measure of "adaptive rationality" (see e.g., Gigerenzer, Todd, & The ABC Research Group, 1999). What are in many ways otherwise closely related programs concerned with heuristics—the Heuristics and Biases program on the one hand, and Gigerenzer and colleagues subsequent search for "simple heuristics that make us smart" (e.g., Todd & Gigerenzer, 2000) on the other—can thus, through a difference in emphasis, come to seemingly strongly conflicting perspectives (see e.g., the exchange Gigerenzer, 1996; Kahneman & Tversky, 1996).

While the later adaptive rationality tradition avoids the term "bias," the words "bias," "error," and "fallacy" figure centrally in the Heuristics and Biases program, and the overwhelming reception of its findings (whatever the original intention) has been as an indictment of human rationality. In the words of Kahneman and Tversky themselves:

> *". . .it soon became apparent that "although errors of judgments are but a method by which some cognitive processes are studied, the method has become a significant part of the message"*
>
> ***(Kahneman & Tversky, 1982, p. 124, and requoted in Kahneman & Tversky, 1996, p. 582)***

In particular, the spectacular success of the program in reaching adjacent disciplines has done much to propagate the notion of human cognition as littered with bias. At the time of writing, Tversky and Kahneman's (1974) Science paper "Judgment under uncertainty: Heuristics and biases" has over 19,000 citations on Google Scholar, and their (1979) paper on decision-making over 28,000, with the majority of these outside of psychology.

This negative assessment of human rationality was perhaps an inevitable side effect of the program's concern with documenting violations of probability theory and decision theory, which themselves have widespread currency as standards of rationality in adjacent disciplines from philosophy to economics. Tversky and Kahneman decidedly took issue with the notion that utility maximization, in particular, provides an empirically adequate descriptive theory of human behavior (e.g., Kahneman & Tversky, 1979). Given that maximization of expected utility effectively defined the "rational man" of economics (see e.g., Simon, 1978), it is unsurprising that a view of people as irrational was the result.

Unlike the 1960s program concerned with "statistical man" just discussed, Tversky and Kahneman focused not on quantitative assessments that sought to identify how closely (or not) human performance matched that of an optimal agent (e.g., measuring degree of conservatism), but rather on *qualitative* violations of probability and decision theory. In many ways, the particular genius of Tversky and Kahneman as experimenters lay in their ability to derive simple problems on which particular patterns of responding would directly indicate normative violations without the need for quantitative modeling.

On the one hand, this makes for simple and compelling demonstrations; on the other hand, however, it does not allow assessment of *how costly* such violations might actually be to people going about their everyday lives. This undoubtedly makes it more tempting to equate "bias" with "irrationality," even though one does not imply the other. As a simple example, consider the conjunction fallacy: assigning a higher probability to the conjunction of two events than to the least probable of the two conjuncts is a simple logical error. The conjunct can be no more probable because, by necessity, the least probable conjunct occurs whenever both events (i.e., the conjunction) occur. One of the most famous fallacies identified by Tversky and Kahneman (see e.g., Tversky & Kahneman, 1983), it implies error by design. Nevertheless, of the many rival explanations for the fallacy (and there are likely many contributing factors, see e.g., Jarvstad & Hahn, 2011 and references therein for

a review), a leading one is that it is the result of a (weighted) averaging strategy for deriving probability estimates (see e.g., Nilsson, Winman, Juslin, & Hansson, 2009). As shown via computer simulation by Juslin, Nilsson, and Winman (2009), such a strategy can provide a remarkably effective combination rule in circumstances where knowledge of the component probabilities is only approximate. Where component estimates are noisy, the multiplicative nature of Bayes' rule means that noise too has a multiplicative effect, an effect that is dampened by additive combination. An additive combination strategy, though normatively incorrect, may thus lead to comparable levels of performance on average, given such noise.

One other well-known bias deserves mention in this context: the tendency for participants to underweight base rates in deriving estimates, a phenomenon labeled "base rate neglect" (Kahneman & Tversky, 1973). It has generated considerable controversy, with literally hundreds of studies investigating the extent to which base rates are underweighted and which circumstances moderate their use (see e.g., Bar-Hillel, 1980; Koehler, 1996).

Not only does an underweighting of base rates fit with an additive combination strategy, and may thus be connected with both the conjunction fallacy and the fact that additive combination may often be a "good" strategy in practical terms (see Juslin et al., 2009), but it also resonates directly with earlier findings discussed under Section 1.1.3 above. In fact, underweighting of base rates replicates earlier findings in bookbag- and pokerchip-like paradigms whereby participants showed sensitivity to the prior probability of hypotheses, but were less sensitive than normatively desirable (see e.g., Green et al., 1964; Wendt, 1969; and for an estimation-only context, e.g., Green, Halbert, & Robinson, 1965). Unlike many studies, those earlier paradigms also allowed assessment of the cost to the participant of deviation from optimal—a cost that in those studies tended to be small (see e.g., Wendt, 1969).

More generally, the case of base rate neglect highlights the need to examine putative biases (as deviations from accuracy) over a broad range of values. This is essential not just for understanding the cost of that deviation but also for the bias' proper scope and interpretation. In the case of low prior probability (e.g., the presence of serious illness such as AIDS, which in the general population has a base rate that is low), underweighting of the base rate means effectively "jumping to conclusions" on the basis of a diagnostic test (such as an AIDS test). Normatively, the actual likelihood of illness given even a high-quality test remains fairly low in light of the low prior probability.

At the other end of the scale, for high prior probabilities, underweighting of base rates means that judgments are not extreme enough. Examining only high *or* low prior probabilities in isolation would lead one to conclude erroneously that people were either too extreme or too hesitant in their judgments, when, in fact, the pattern of responding is indicative of a more general "conservatism," that is, sensitivity to normatively relevant factors, but by not enough.

In the decades since, judgment and decision-making research has chipped away at the discrepancies between normative and descriptive highlighted by the Heuristics and Biases program in a number of distinct ways (though typically with considerably lower profile than the original negative news, see e.g., Christensen-Szalinski & Beach, 1984). In particular, it has been argued that seeming errors may stem from divergent construals of the task by experimenters and participants (e.g., Hilton, 1995; Schwarz, 1996). There have also been arguments over normative evaluations of the tasks (e.g., Gigerenzer, 1991; Koehler, 1996), both in judgment and decision-making and in the context of investigations of human rationality in adjacent fields such as logical reasoning (e.g., Oaksford & Chater, 1994, 2007; Wason, 1968).

It has also been demonstrated that the biases in question are far from universal (e.g., Stanovich & West, 2000). In studies of the conjunction fallacy, for example, there is typically a proportion of participants who do not commit the fallacy (see e.g., Jarvstad & Hahn, 2011). This may be taken to be indicative of interesting facts about cognitive architecture (such as the existence of multiple cognitive "systems" capable of generating responses, see e.g., Kahneman, 2000). However, it may also be taken to undermine the very project. Rather than providing evidence of systematic and pervasive irrationality, the existence of stable individual differences in susceptibility to bias could be taken to imply "that the vast literature on heuristics and biases may embody little more than a collection of brain teasers that most people get wrong but that a few people—without tutoring and despite everything—manage to get right" (Funder, 2000, p. 674). Viewed from that perspective, this tradition of research does not reveal systematic irrationality, but "variations in the ability to answer difficult questions," where "some questions are so difficult that only very smart people get them right"—a state of affairs that is intrinsically no more interesting and no more informative of the nature of human cognition than that SATs (scholastic aptitude tests administered to students in the US) contain questions that most students will get wrong (Funder, 2000).

1.1.5 Social Psychology

While the Heuristics and Biases program came to dominate cognitive psychology (more specifically, judgment and decision-making research), its impact in social psychology was less strong. There too, it is perceived to have become increasingly influential (see e.g., Krueger & Funder, 2004 for discussion of this point), but social psychology contains much distinct work of its own concerned with bias and error (with often seemingly little distinction between the two, see also Kruglanski & Ajzen, 1983). In fact, it has attracted high-profile critiques of its unduly "negative" focus. Krueger and Funder (2004) lament that

> *. . .social psychology is badly out of balance, that research on misbehavior has crowded out research on positive behaviors, that research on cognitive errors has crowded out research on the sources of cognitive accomplishment, and that the theoretical development of social psychology has become self-limiting.*
>
> **(Krueger & Funder, 2004, p. 322)**

As a consequence, social psychology, in Krueger and Funder's perception, has accumulated a long list of putative, often contradictory, biases (e.g., false consensus effect and false uniqueness effect, see e.g., Table 1 of Krueger & Funder, 2004), a list that continues to grow as variants of old biases are rediscovered with new names. This, in their view, has led to a warped, unduly negative, overall assessment of human competence, while providing little insight into underlying mental processes.

The wave of research into errors and biases (according to e.g., Funder, 1987; Kenny & Albright, 1987) is seen in part as a response to the demise of early research into the accuracy of interpersonal perception within social and personality psychology that was brought about by devastating methodological critiques of standard methods (in particular, critiques by Cronbach, 1955; Gage & Cronbach, 1955; Gage, Leavitt, & Stone, 1956). In the context of social judgment, "truth" is hard to come by: if someone perceives someone to be "friendly" on the basis of their interactions thus far, it is hard to establish the criterion value against which their accuracy might be assessed. In light of these difficulties, the preferred method in early research on the accuracy of interpersonal judgments was to get members of a group to provide judgments about each other and then to evaluate their accuracy in terms of how well they agreed (see e.g., Dymond, 1949, 1950). Cronbach and Gage's critiques demonstrated exactly how difficult such data were to interpret, more or less bringing research in this tradition to a halt. Subsequent research on social judgment sought to "bypass

the accuracy problem" by turning to the study of errors and biases brought about by the use of heuristics, by (unwarranted) implicit assumptions, and by "egocentric orientation" (Kenny & Albright, 1987).

In so doing, the error and bias tradition within social judgment proceeded by making "normative standards" inherent in the experimental manipulation itself. For example, an (erroneous) judgmental tendency to overascribe behavior to enduring dispositions (as opposed to situational factors) was inferred from experimental paradigms in which dispositional information is (supposedly) rendered irrelevant by experimental design. In Jones and Harris (1967) classic study of attribution, participants were shown essays favoring Fidel Castro that were purportedly written by people who had no choice in writing a pro-Castro piece. Participants nevertheless showed a tendency to view pro-Castro essays as reflecting "true" pro-Castro positions on the part of the authors.

In this way, differential responses to experimental materials become evidence of error and bias. In particular, evidence for motivated distortions of evidence have been sought in this way. For example, Lord et al. (1979) famously demonstrated "biased assimilation" in this manner. In their study, participants were presented with mixed evidence on the effectiveness of capital punishment in deterring crime. Each participant read two (experimenter designed) journal articles, one purporting to show effectiveness and the other purporting to show ineffectiveness. Participants rated the report that agreed with their prior opinion as "more convincing," and more readily found flaws in the reports that went against it. Moreover, the effect of each report on the participant's subsequent beliefs was stronger when the report agreed with their prior self-assessment as proponents or opponents of capital punishment. In other words, participants' beliefs became more polarized by conflicting evidence that, if anything, should have made them less sure of their beliefs.

To the extent that there is a general rationality principle that might be articulated for such cases, it is what Baron (2008) calls the "neutral evidence principle": "Neutral evidence should not strengthen belief," that is evidence that is equally consistent with a belief and its converse, such as mixed evidence, should not alter our beliefs. This neutral evidence principle is violated when ambiguous evidence is interpreted as supporting a favored belief.

In many ways, the notion of "bias" operative here is thus close to the lay meaning of bias as "lack of impartiality." However, it is typically assumed that such a bias will also have systematic negative effects on the accuracy of our beliefs (see e.g., Baron, 2008).

Other methods for examining bias in social judgment make use of comparative ratings. In the many studies concerned with self-enhancement biases, for example, the true level of a participant's skill (e.g., Svenson, 1981), or risk of experiencing an adverse life event (e.g., Weinstein, 1980), etc., is unknown. In these circumstances, bias is ascertained via a comparison between multiple quantities, such as self versus other perception, or self versus average and so on. The logic here is that while it may be impossible to say whether a given individual is a "better-than-average" driver or not, (sufficiently large) groups of individuals rating their driving skills should come to match average values. Intuitive as that may seem, the application of formal models has shown such reference point dependent evaluations to be prone to statistical artifacts, in particular regression artifacts (see e.g., Fiedler & Krueger, 2011).

In general, it may be said that the types of bias observed within social psychology are, if anything, seen as even more "irrational" than those observed by cognitive psychologists in the context of judgment and decision-making:

> *Motivational biases are characterized by a tendency to form and hold beliefs that serve the individual's needs and desires. Individuals are said to avoid drawing inferences they would find distasteful, and to prefer inferences that are pleasing or need-congruent. Being dependent on the momentary salience of different needs, such motivational influences could presumably yield judgmental biases and errors. Even in the absence of motivated distortions, human judgments are assumed subject to biases of a more cognitive nature. Unlike motivational biases that are presumed to constitute largely irrational tendencies, cognitive biases are said to originate in the limitations of otherwise reasonable information-processing strategies.*
>
> ***(Kruglanski & Ajzen, 1983, p. 4)***

In other words, putative motivational biases stem from a tendency to engage in "wishful thinking" in order to maintain self-serving motives such as the need for "self-enhancement" or "effective control," whereas cognitive biases are mere side effects from the use of suboptimal judgment heuristics or strategy. From the perspective of the biased agent, motivational and cognitive biases thus differ in fundamental ways: for the former, the bias is, in a sense, the goal; for the latter, it is an (undesirable) by-product of a system that is otherwise striving for accuracy.

Both types of bias should violate normative models of judgment such as Bayes' rule but, unlike the Heuristics and Biases tradition, social psychological research typically examined bias (and error) independently of normative framework. Kruglanski and Ajzen (1983) noted 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 & Ajzen, 1983 p. 2)***

This state of affairs has largely remained, and means that in many social psychological studies concerned with bias, there is simply no clearly articulated standard of rationality (see also, Griffin, Gonzalez, & Varey, 2001; Krueger & Funder, 2004).

Earlier research in social psychology had seen some reference to Bayesian belief revision; for example, Ajzen and Fishbein (1975) argued that a wealth of different findings on causal attribution might be understood (and thus unified) from a Bayesian perspective. In other words, experimental variations probing factors influencing causal attribution can typically be recast as manipulations that affect the diagnosticity (and hence evidential value) of the information given to participants. So their responses may be understood as tracking that information in a process of subjective belief revision that approximates the Bayesian norm in the same way that participants respond in broadly qualitatively appropriate ways to probabilistically relevant factors in bookbag and pokerchip paradigms. Such reference to the Bayesian framework, however, is the exception rather than the rule.

Moreover, even those social psychologists who have taken issue with bias focused research (e.g., Funder, 1995; Kruglanski, 1989; Kruglanski & Ajzen, 1983; Krueger & Funder, 2004) and have argued strongly for a research focus on accuracy using tasks where accuracy can be meaningfully defined, have tended to express some skepticism towards the use of normative models on the grounds that there is debate about normative standards of rationality, and there may thus be rival "norms" (see also, Elqayam & Evans, 2011).

Such debate has several sources. Even for tasks such as probability judgment for which there is considerable consensus about norms, applying these to particular experimental tasks and questions may be less than straightforward and much of the critical debate surrounding the heuristics and biases tradition originates here (see e.g., Gigerenzer, 1991; Koehler, 1996). In other cases, debate about normative standards is more fundamental, with ongoing debate about how best to conduct the inference in question, that is, norms themselves. For example, much research on causal attribution was closely related to particular, classical (frequentist) statistics such as analysis of variance (see e.g., Kelley & Michela, 1980), for which there are now rival

statistics. In other cases, general assumptions about how science should proceed provided the role model. Neither the philosophy of science nor epistemology, however, are "completed," and both have been subject to considerable development and debate, in particular a move from an emphasis on deduction as found in Popperian falsification (Popper, 1959) to a more recent emphasis on Bayesian probability (see e.g., Howson & Urbach, 1996). In the meantime, social psychologists themselves have sought to formulate their own perspectives of "lay epistemology" (e.g., Kruglanski & Ajzen, 1983).

At the same time, researchers within social (and personality) psychology, have considered broader conceptions of "truth" and consequently "accuracy" (see e.g., Funder, 1995; Kruglanski, 1989). These encompass not just accuracy as "correspondence between a judgment and a criterion" (in parallel to correspondence theories of truth, see e.g., Funder, 1987; Hastie & Rasinski, 1988; Kenny & Albright, 1987), but also a constructivist perspective that views "accuracy as interpersonal agreement between judges" (e.g., Kruglanski, 1989) and a conceptualization of the accuracy of a judgment in terms of its adaptive value (in keeping with a pragmatic notion of truth, see e.g., McArthur & Baron, 1983; Swann, 1984).

There are thus multiple reasons why social psychologists have not always viewed biases as ultimately "bad." Self-enhancement biases have been taken to provide "cognitive illusion" that promote well-being and mental health (e.g., Taylor & Brown, 1988), biases have been argued to (sometimes) promote accuracy in person perception (e.g., Funder, 1995), and they have been argued to reflect evolutionary adaptations to asymmetric costs of errors (e.g., "error management theory," Haselton & Buss, 2000; Haselton & Funder, 2006). More generally, a focus on adaptive consequences may give rise to a terminological distinction between error and bias itself. McArthur and Baron's (1983) ecological perspective suggests that

> *. . .bias is different from error: Bias is simply a matter of selective attention and action, and whether a given bias leads to error in adaptive behavior is an empirical, not a logical, problem. (p. 230)*

Likewise, a focus on real-world outcomes led Funder (1987) to distinguish between "errors" and "mistakes":

> *. . .An error is a judgment of an experimental stimulus that departs from a model of the judgment process. If this model is normative, then the error can be said to represent an incorrect judgment. A mistake, by contrast, is an incorrect judgment of a real-world stimulus and therefore more difficult to determine.*

> *Although errors can be highly informative about the process of judgment in general, they are not necessarily relevant to the content or accuracy of particular judgments, because errors in a laboratory may not be mistakes with respect to a broader, more realistic frame of reference and the processes that produce such errors might lead to correct decisions and adaptive outcomes in real life. (p. 75)*

From here, it no longer seems surprising that a recent methodological proposal for the study of social judgment by West and Kenny (2011) employs a notion of "bias" that encompasses *any evidence* (other than simply the truth itself) that may be used by an agent to infer some quantity to be judged.

1.1.6 Summary

It seems fair to describe the use of the term "bias" within psychological research as varied, at times encompassing almost polar opposites: the term has been used to denote both systematic deviations from accuracy and mere error, it has been taken to reflect both "outcome" and process, a side effect and a goal, and bias has been viewed variously as obviously irrational, as rational, or neither.

In a sense, any terminology is fine as long as it is clear. However, terminological confusion tends to obscure important empirical issues. Our goal in this chapter is to lend some precision particularly to what has and has not been shown in the domain of "motivated cognitions." For this, it is useful to provide some indication of more formal notions of "bias" within statistics. We discuss these next, before returning to a more in depth look at wishful thinking and motivated reasoning.

1.2. The Notion of Bias in Statistics

1.2.1 Bias as Expected Deviation

In statistics (and related disciplines such as machine learning) the term "bias" refers to (expected) systematic deviation (see e.g., Bolstad, 2004). If, for example, we are trying to estimate a proportion, such as the proportion of the population who will contract a particular disease, on the basis of a sample, then the bias of an estimator (a statistic for estimating that proportion) is the difference between the expected value of that estimator and the true population proportion:

$$\text{bias}(\text{estimator}) = E(\text{estimator}) - \text{True Population Value}$$

in other words, the difference between the true proportion and the average value of the estimator (over samples).

Similarly, consider a case in which we try to estimate a function on the basis of a (finite) sample of data, so that we may generalise to other, as yet unseen, values. Our predictor is said to be "biased" if the average value of our predictor is different from the true value (or where there is noise, from its expectation).

"Bias" may thus intuitively seem like a "bad thing." However, the situation is more complicated. If, as is common, we evaluate our accuracy in terms of mean squared error (the average squared distance of the estimator from the "true value") then

$$\mathrm{MSE} = \mathrm{bias}(\mathrm{estimator})^2 + \mathrm{variance}(\mathrm{estimator})$$

Thus, if it has smaller variance, a biased estimator may be more accurate, on average, than an unbiased estimator and provide a value that is closer to the truth.

As an illustration, we consider an example from the current debate in psychology on whether to use classical, frequentist statistics or "Bayesian" counterparts (e.g., Kruschke, 2010; but see also already Edwards, Lindman, & Savage, 1963). The standard frequentist measure for estimating a population proportion, such as those who contract a particular disease, is the sample mean: the proportion of diseased within our sample. The sample mean is an unbiased estimator.

Alternatively one might seek to estimate that same proportion in a different way. We think of the true proportion as a particular value from a distribution of possible values (ranging from 0 to 1) and calculate a posterior distribution in light of our sample, taking the mean of that posterior to be our estimator. In this case, we need to choose a prior distribution that is combined with our evidence via Bayes' theorem to calculate that posterior. A standard choice (but it is a choice!) would be a beta distribution as a prior with values that give a uniform distribution over all values between 0 and 1 (i.e., beta(1,1)), reflecting a lack of any knowledge that would make some proportions more or less likely *a priori*. This measure is not unbiased (and typically Bayesian statistics are not), yet its average mean squared error (over the range of possible true values of the population proportion) is lower than that of the unbiased sample mean. As seen in Fig. 2.1, which shows both bias and variance components and resulting MSE (for formulae used in calculation, see e.g., Bolstad, 2004) for sample sizes of 5, 10, and 20, the posterior mean outperforms the sample mean not just in particular "lucky" cases, but does better for most (though not all) possible population

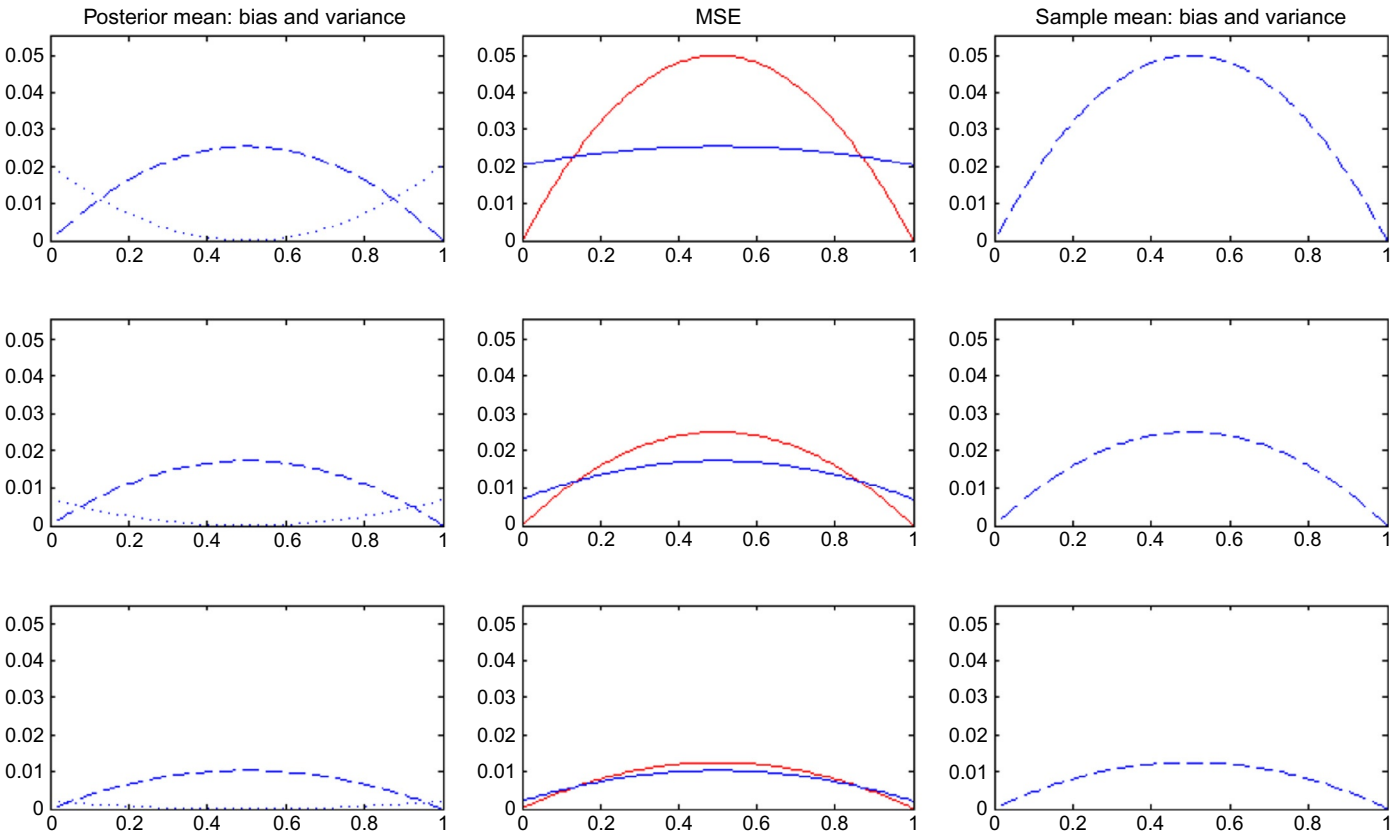

Figure 2.1 Bias, variance, and mean squared error (MSE) for two estimators of sample proportion, the sample mean, and the posterior mean of a beta distribution with uniform prior. Each row corresponds to a different sample size, n: top row $n = 5$, middle row $n = 10$, bottom row $n = 20$. The left hand column shows the squared bias (dotted line) and variance (dashed line) of the posterior mean across range of possible population proportions (x-axis). The right hand column shows bias (always 0) and variance for the sample mean. The middle column shows the MSE error of both predictors (sample mean, grey line (red in online version); posterior mean, black line (blue in online version)) which is the sum of squared bias and variance. As can be seen, MSE is lower for the posterior mean across most of the range, and always lower on average.

proportions, and does better on average. If accuracy is our goal (as seems reasonable in this context), we may consequently be better off with a "biased" estimator.

Moreover, it may not be possible to minimize *both* bias and variance. As Geman, Bienenstock, and Doursat (1992) show for the case of generalization, decreasing bias may come at the expense of increasing variance, and vice versa—a phenomenon they refer to as the "bias/variance dilemma."

Consider the case of a feed-forward neural network trained by back-propagation (e.g., Ripley, 1996). Such networks perform a type of (nonparametric) regression. A small network with a very limited number of hidden units is likely to be quite biased, as the range of functions that can be captured exactly over the possible hidden unit weights will be restricted. Increasing the number of hidden units will reduce bias, but increasing the number of parameters means that the variance increases: the network will (over)fit the training data meaning that generalization predictions will be tied too closely to the specific characteristics of the training sample and will vary widely with variation in that sample as opposed to robustly approximating the underlying function. Figure 2.2 shows an example reproduced from Geman et al. (1992) involving neural networks learning to classify handwritten digits. The figure shows total error on a test set of 600 images of handwritten digits, after training on an independent set of 200 images. Bias and variance are approximated by averaging over repeated simulations of networks with different numbers of hidden units. As can be seen, small networks show high bias and low variance, large networks, low bias and high variance.

At the one extreme, a completely biased learner is oblivious to the data; at the other, the learner is so sensitive to the characteristics of the particular sample that no meaningful generalization to new instances occurs. For fixed samples, optimal learning requires a balance between constraints on the learner, which introduce bias, and variance that arises as a result of sensitivity to the data. The best performance will be obtained by a learner that has a bias suitable to the problem at hand.[2]

It should be noted that the exact relationship between bias and variance depends on the type of problem and the measure of success that is appropriate (i.e., the loss function or scoring rule, see e.g., Domingos, 2000; Wolpert, 1997). However, the above examples suffice to show that, in the context of statistics and machine learning, "bias" does not equal "bad."

[2] This trade-off also underlies Gigerenzer and colleagues arguments for adaptive heuristics (see e.g., Gigerenzer, 2008).

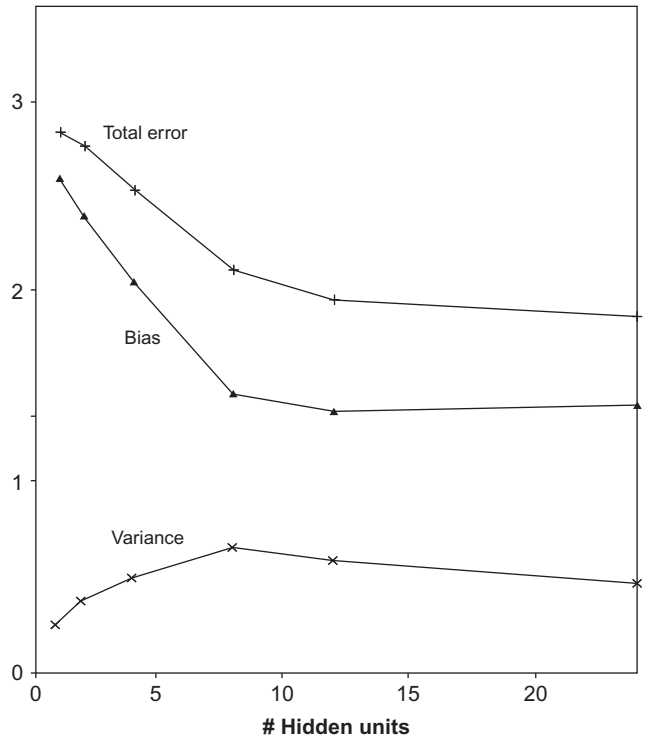

Figure 2.2 Reproduced from Geman et al. (1992). The *x*-axis displays the number of hidden units in the network, the *y*-axis the mean squared error across the test set. Each entry is the result of 50 network trials except for the last entry (24 hidden units) which is based on 10 trials only. In each case, network learning (i.e., the number of sweeps through the training set) was continued to the optimal point, that is, the point that minimized total error.

1.2.2 Signal Detection Theory

The discussion of bias in statistics so far has focused on generating estimates, but we often also need to make decisions. Indeed, even assigning an item to a discrete category on the basis of the evidence involves a decision of sorts. In the context of decisions, "bias" corresponds to a preference for one of several decision outcomes. As we saw in the everyday use of the word, a bias in the context of choices is manifest in selections that are based on something above and beyond "intrinsic merits" of the options.

A familiar use of the term bias in this decision-based sense arises in signal detection theory (SDT), a popular technique for modeling empirical decision processes. SDT was originally derived from statistical decision theory (Berger, 1985; Wald, 1950) in order to relate choice behavior to a

psychological decision space for a wide range of underlying tasks such as discrimination, recognition, or classification (see e.g., Green & Swets, 1966; Swets, 1964; see also Pastore, Crawley, Berens, & Skelly, 2003, for a history). SDT has found application in psychology not just for the study of perceptual processes, but also memory, or medical diagnosis, and has also seen increasing application in adjacent fields such as forecasting (see e.g., Swets, Dawes, & Monahan, 2000, for references).

Statistical decision theory, in general, seeks to define optimal decision strategies in situations where evidence itself is noisy or uncertain. An optimal approach involves evaluating the likelihood that an observed value of that evidence has arisen from each of a range of alternative hypotheses that are being considered. Optimal decisions should reflect those likelihoods, but should also take into account potential asymmetries in costs and benefits (where they exist).

SDT is an application of statistical decision theory to the modeling of human decision behavior, providing a set of measures that allow the decomposition of performance into distinct contributing components. Specifically, it is assumed that the decision-maker aggregates evidence and evaluates the likelihood of obtaining that evidence under each of two alternative hypotheses (e.g., "signal present, no signal present," "word/nonword," "old item/new item," though generalizations to multiple hypotheses have also been derived, see e.g., DeCarlo, 2012). The likelihood comparisons can be represented along a single underlying dimension representing, for example, the ratio of the contrasted likelihoods—the so-called likelihood ratio (LHR)[3]—that is, the probability of the evidence obtained given that Hypothesis 1 is true, $P(e|H1)$, divided by the probability of obtaining that evidence if Hypothesis 2 were true, $P(e|H2)$ (see Pastore et al., 2003). This provides an underlying measure of "accumulated evidence."

In order to select a response ("*H*1" or "*H*2," "old item," or "new item," etc.), the decision-maker must select a threshold on this underlying continuous dimension, whereby values above the threshold receive one response, and values below receive the other response. There are thus two factors that will affect overall performance: (1) how well the evidence evaluation discriminates between the two alternative hypotheses and (2) where the decision threshold is placed. The literature contains a wealth of terms to

[3] The underlying dimension may also be a monotonic transformation of the LHR (see Pastore et al., 2003). By contrast, the widespread characterization of the underlying dimension as reflecting "a single sensory continuum" as opposed to a measure of accumulated evidence is incorrect (see Pastore et al., 2003).

refer to each of these. To avoid confusion with other concepts in this chapter, we will refer to (1) as "discrimination ability" and (2) as the "decision criterion."[4]

The relationship between these two components in determining overall performance is illustrated by the so-called receiver operating curve (ROC), see Fig. 2.3 for an example. An ROC plot shows the impact of shifting the

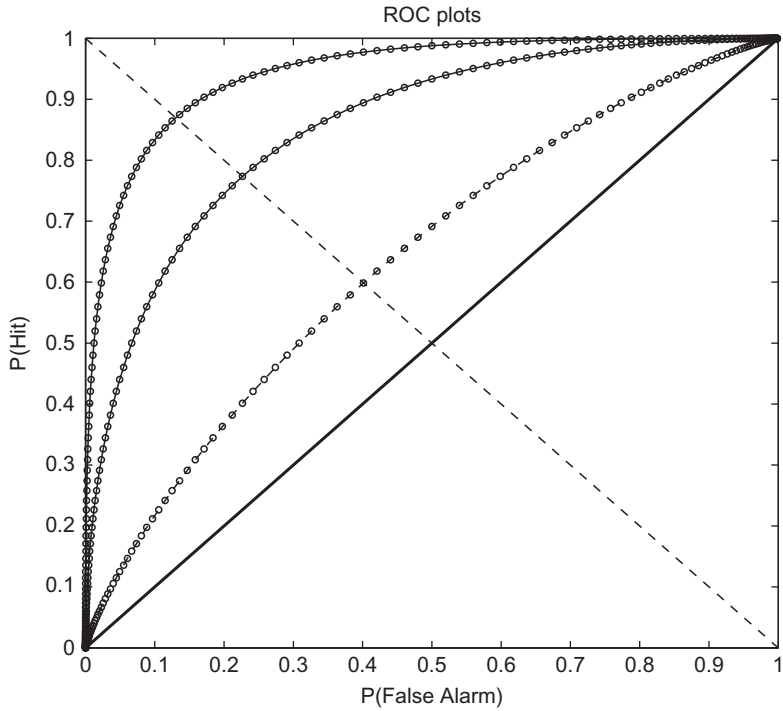

Figure 2.3 Illustrative ROC curves. On the *x*-axis is the probability of a false alarm (false positive), on the *y*-axis is the probability of a "hit." The positive diagonal (full line) represents inability to discriminate between the two alternatives. The different curves represent different levels of discrimination ability, with discrimination ability increasing with the area under the curve. The negative diagonal (dashed line) indicates an unbiased decision criterion. The intersection of that diagonal with each of the ROC curves indicates the hit rate and false-positive rate for an unbiased response selection at that level of discrimination ability. Alternative (biased) decision criteria are represented by circles along a given curve. In other words, picking different decision criteria on the part of the responder corresponds to moving along a given curve, and thus systematically shifting the hit and false alarm rates.

[4] Discrimination ability is also referred to as "accuracy" (see e.g., Pastore et al., 2003) or "sensitivity" (see e.g., Stanislaw & Todorov, 1999), the criterion is also referred to as "response bias," or "bias" (see e.g., Pastore et al., 2003).

decision criterion on patterns of responding. On the *y*-axis is plotted the probability of a "hit" (a correct identification of hypothesis *H*1), and on the *y*-axis the "false alarms" (the probability of incorrectly responding *H*1, when it is *H*2 that is true). A given level of discrimination ability marks out a curve in this space, and different points along this curve correspond to different values of the decision criterion that could be chosen (i.e., the slope of the curve at that point is equal to the LHR, see Swets et al., 2000). Because discrimination ability is imperfect, adopting a more liberal decision criterion—that is, requiring less compelling evidence in favor of hypothesis *H*1—will not only affect the hit rate but also the false-positive rate. Setting a less stringent criterion in evaluating a mammogram will lead to more referrals for further examination. This will not only increase the hit rate (detecting cancer) but also generate more false positives (giving rise to what turns out to be needless intervention). Where a decision-maker best sets the criterion will consequently depend in part on overall goals.

The decision criterion that simply always selects the hypothesis with the higher likelihood[5] is *unbiased*. In Fig. 2.3, it is represented by the negative diagonal indicated by the dotted line. However, once again "bias" does not equal "bad." The unbiased decision criterion only maximizes accuracy (i.e., the proportion of correct responses[6]), when the prior probabilities of both hypotheses are the same, that is, the base rates of the two alternatives are equal. If cancer is rare, then an unbiased decision criterion will lead to many false positives.

Moreover, the unbiased decision criterion does not factor in costs and benefits that might be associated with the selection of different alternatives. The benefits of the correct decision outcome(s) may far outweigh the costs of the incorrect selections, or vice versa.

Given that the decision-maker does not know the true state of the world, the optimal decision criterion must be based on expectations. As with other decisions involving uncertainty, the optimal decision policy should maximize the *expected value* of the response selection. According to Swets et al. (2000) it was first derived by Peterson, Birdsall, and Fox (1954) what the exact relationship between base rates, costs, and benefits for this is. The expected value for choosing one of the two response alternatives is defined as follows. The decision-maker is seeking to determine which of two hypotheses, *H*1 or *H*2, is correct (e.g., a medical condition is present or

[5] That is, the decision criterion = 1, and the hypothesis in the numerator is chosen whenever the LHR is >1, and the hypothesis in the denominator when it is less than 1.

[6] Accuracy in this sense is equal to (Hits + Correct Rejections)/(Hits + Correct Rejections + False Positives + Incorrect Rejections), where a correct rejection is a response of "not *H*1" (i.e., *H*2) when *H*2 is true.

absent, the word on a given memory trial is "old" or "new," etc.). D_{H1} represents the selection of $H1$ as the chosen response, and the same for D_{H2}; C and B represent "costs" and "benefits" so that $B(D_{H1}\ \&\ H1)$ represents the benefits of choosing $H1$ as the response when in fact $H1$ turns out to be true (i.e., a "hit"), and $C(D_{H1}\ \&\ H2)$ represents the costs of a "false positive," that is, responding $H1$ when $H2$ is in fact the case. $P(H1)$ and $P(H2)$ represent the prior probabilities (base rates) of $H1$ and $H2$. The optimal decision criterion $C(\text{optimal})$ is then given as[7]:

$$C(\text{optimal}) = \frac{P(H1)}{P(H2)} \times \frac{B(D_{H2}\ \&\ H2) + C(D_{H1}\ \&\ H2)}{B(D_{H1}\ \&\ H1) + C(D_{H2}\ \&\ H1)} \tag{2.1}$$

As can be seen from this formula, if all benefits and costs are considered equal (i.e., their ratio is 1.0), they play no role in determining the expected value of a selected response; in this case, it is the prior probabilities alone that determine the optimal threshold (Swets et al., 2000).

In summary, according to the normative prescriptions of decision theory, it *is rational to be biased* in responding whenever the base rates of the two alternatives are not equal and/or the relevant cost/benefit ratios are unequal. In these cases, which will be numerous in the real world, it would be irrational to adopt an unbiased decision criterion in the sense that it would lead, on average, to less good outcomes for the decision-maker. Ignoring base rates will reduce accuracy, and ignoring costs and benefits will mean missing out in terms of overall consequences (see Harris & Osman, 2012, on the importance of such parameters in understanding the status of the illusion of control as an (ir)rational bias).

For a decision criterion to be irrational, the bias of the decision-maker must deviate from Eq. (2.1). It is, of course, entirely possible that people's biases do deviate and that they are both irrational and maladaptive in this sense. However, base rates (and beliefs about them), costs, and benefits are rarely explicitly assessed in actual experiments. In fact, much of the appeal of SDT stems from the fact that it provides statistical procedures for estimating discrimination ability and decision criterion from empirical data in circumstances where underlying probability distributions governing the decision are not known (see Pastore et al., 2003). Experiments may seek to control base rates and costs and benefits associated with correct and

[7] For C as the slope at any given point along the ROC curve when the underlying continuum represents the LHR see Swets et al. (2000). For other conceptualizations of the underlying "evidence" dimension in the SDT model, other relationships may apply.

incorrect responding by experimental design; however, this still requires that the participant's perceptions of the situation match those of the experimenter and this cannot simply be taken for granted. It has been a recurring theme of critiques of experiments supposedly highlighting human irrationality that their results depend on systematic (and misleading) violations of participant expectations and that when discrepancies are reduced, so are the deviations from normative, rational responding (see e.g., Hilton, 1996; Schwarz, 1996 for discussion and specific examples).

1.3. Implications

Several general points emerge from the preceding general overview of research concerned with "bias." The first of these is that "bias" is neither necessarily irrational nor bad in any wider sense. "Bias" in the sense of response bias may be *optimal* when costs of hits, false positives and correct and incorrect rejections are unequal (as is capitalized on in e.g., error management theory; Haselton & Buss, 2000). In the remainder of this chapter, we concern ourselves only with accuracy goals. When a bias compromises accuracy goals, it makes sense to consider whether there may be secondary justifications for it. Where a bias does not compromise accuracy, there is no need to look for further justification (and, in fact, that further justification will be baseless), nor need one look for adaptive rationales where the mere existence of bias is not even clearly established. A focus on accuracy thus seems a necessary first step. Here, a bias may be desirable even where it is only accuracy one cares about because "response bias" (a biased decision criterion) is a consequence of optimal responding in the case of unequal priors. Moreover, the desirability of bias in estimators is generally subject to trade-offs.

This has several implications for establishing the presence of costly "bias." In order to show that an estimation process is biased in a way that will compromise the accuracy of people's belief systems, one needs to show more than that it is sometimes wrong. Rather,

1. "bias" must be understood as a property of an estimator that holds for an *expectation*, that is on average
2. this expectation must be calculated over a broad range of values in order to allow meaningful evaluation, that is, it needs to be shown that the estimator is *systematically wrong* across different contexts
3. and, finally it needs to be shown that it is wrong *at a cost* (in first instance an accuracy cost, though other costs are of course relevant in principle)

It may be argued that most research on "bias" falls short in one or more of these respects. Research on conservatism in the bookbag and pokerchip tradition has gone furthest at meeting these requirements. In Kahneman and Tversky's Heuristics and Biases program, the issue of accuracy costs (3) is typically unaddressed. In fact, it may be argued that many of their violations of decision-theoretic norms, for example, have been obtained in contexts where the two options presented for choice differ so minimally in expected value that such violations come at virtually no cost (see for recent examples, Jarvstad, Hahn, Rushton, & Warren, 2013; Jarvstad, Hahn, Warren, & Rushton, 2014; the general issue is discussed in detail by Winterfeldt and Edwards (1982) under the header of "flat maxima"). Furthermore, accuracy costs may also have implications for (2), that is the systematicity and scope of the bias, because it seems possible that the application of heuristics may be confined to cases where there is little cost in getting things wrong and that optimal strategies are applied elsewhere (for evidence to this effect see also, Brandstätter, Gigerenzer, & Hertwig, 2006).

It should be equally clear that much of the social psychological research discussed above already fails to meet requirement (1): a demonstration that participants process a few pieces of information in what, by experimental design, seems like a "biased" way does not even allow evaluation of the average impact of such behavior (if indeed it generalizes beyond the confines of that experiment). In this case, it is simply assumed that the behavior in question extends in ways that (1)–(3) are met. Such extrapolation, however, is perilous and we seek to demonstrate the frailties of such inference in the remainder, drawing on examples from research findings that have been taken as evidence of "motivated reasoning." In so doing, we show why such extrapolation invariably requires reference to optimal (normative) models.

Specifically, the remainder of the chapter will provide detailed examination of criteria (1)–(3) in motivated reasoning research, in particular in the context of wishful thinking, confirmation bias in evidence selection, biased assimilation of evidence, and the evidence neutrality principle.

2. WHEN IS A BIAS A BIAS?

2.1. Understanding Bias: Scope, Sources, and Systematicity

We begin our example-based discussion with a very general bias which, if robust, would provide direct evidence of motivated reasoning, namely "wishful thinking." Under this header, researchers (mostly in the field of

judgment and decision-making) group evidence for systematic overestimation in the perceived probability of outcomes that are somehow viewed as desirable, as opposed to undesirable.

In actual fact, robust evidence for such a biasing effect of utilities or values on judgments of probability has been hard to come by, despite decades of interest, and the phenomenon has been the dubbed "the elusive wishful thinking effect" (Bar-Hillel & Budescu, 1995). Research on wishful thinking in probability judgment has generally failed to find evidence of wishful thinking under well-controlled laboratory conditions (see for results and critical discussion of previous research, e.g., Bar-Hillel & Budescu, 1995; Bar-Hillel, Budescu, & Amar, 2008; Harris, Corner, & Hahn, 2009). There have been observations of the "wishful thinking effect" outside the laboratory (e.g., Babad & Katz, 1991; Simmons & Massey, 2012). These, however, seem well explained as "an unbiased evaluation of a biased body of evidence" (Bar-Hillel & Budescu, 1995, p. 100, see also Gordon, Franklin, & Beck, 2005; Kunda, 1990; Morlock, 1967; Radzevick & Moore, 2008; Slovic, 1966). For example, Bar-Hillel et al. (2008) observed potential evidence of wishful thinking in the prediction of results in the 2002 and 2006 football World Cups. However, further investigation showed that these results were more parsimoniously explained as resulting from a salience effect than from a "magical wishful thinking effect" (Bar-Hillel et al., 2008, p. 282). Specifically, they seemed to stem from a shift in focus that biases information accumulation and not from any direct biasing effect of desirability. Hence, there is little evidence for a general "I wish for, therefore I believe. . ." relationship (Bar-Hillel et al., 2008, p. 283) between desirability and estimates of probability. Krizan and Windschitl's (2007) review concludes that while there are circumstances that can lead to desirability indirectly influencing probability estimates through a number of potential mediators, there is little evidence that desirability directly biases estimates of probability.

What is at issue here is the systematicity of the putative bias—the difficulty of establishing the presence of the bias across a range circumstances. The range of contexts in which a systematic deviation between true and estimated value will be observed depends directly on the underlying process that gives rise to that mismatch. Bar-Hillel and Budescu's (1995) contrast between "an unbiased evaluation of a biased body of evidence" and a "magical wishful thinking effect" reflects Macdougall's (1906) distinction between "primary" and "secondary bias," namely a contrast between selective information uptake and a judgmental distortion of information so acquired.

Both may, in principle, give rise to systematic deviations between (expected) estimate and true value; however, judgmental distortion is more pernicious in that it will produce the expected deviation much more reliably. This follows readily from the fact that selective uptake of information cannot, by definition, guarantee the *content* of that information. Selectivity in where to look may have some degree of correlation with content, and hence lead to a selective (and truth distorting) evidential basis. However, that relationship must be less than perfect, simply because information uptake on the basis of the content of the evidence itself would require processing of that content, and thus fall under "judgmental distortion" (as a decision to neglect information already "acquired").

In fact, selective attention to some sources over others can have a systematic effect on information content *only* where sources and content are systematically aligned and can be identified in advance.

Nevertheless, selectivity in search may lead to measurable decrements in accuracy if it means that information search does not maximize the expected value of information. In other words, even though a search strategy cannot guarantee the content of my beliefs (because there is no way of knowing whether the evidence, once obtained, will actually favor or disfavor my preferred hypothesis), my beliefs may systematically be less accurate because I have not obtained the evidence that could be expected to be most informative.

This is the idea behind Wason's (1960) confirmation bias. Though the term "confirmation bias," as noted, now includes phenomena that do not concern information search (see earlier, Fischhoff & Beyth-Marom, 1983), but rather information evaluation (e.g., a potential tendency to reinterpret or discredit information that goes against a current belief, e.g., Lord et al., 1979; Nisbett & Ross, 1980; Ross & Lepper, 1980), Wason's original meaning concerns information acquisition. In that context, Klayman and Ha (1989) point out that it is essential to distinguish two notions of "seeking confirmation":

1. examining instances most expected to verify, rather than falsify, the (currently) preferred hypothesis.
2. examining instances that—if the currently preferred hypothesis is true—will fall under its scope.

Concerning the first sense, "disconfirmation" is more powerful in deterministic environments, because a single counter-example will rule out a hypothesis, whereas confirming evidence is not sufficient to establish the truth of an inductively derived hypothesis. This logic, which underlies Popper's (1959)

call for falsificationist strategies in science, however, does not apply in probabilistic environments where feedback is noisy. Here, the optimal strategy is to select information so as to maximize its expected value (see e.g., Edwards, 1965; and on the general issue in the context of science, see e.g., Howson & Urbach, 1996). In neither the deterministic nor the probabilistic case, however, is it necessarily wrong to seek confirmation in the second sense—that is, in the form of a positive test strategy. Though such a strategy led to poorer performance in Wason's (1960) study this is not generally the case and, for many (and realistic) hypotheses and environments, a positive test strategy is, in fact, more effective (see also, Oaksford & Chater, 1994).[8] This both limits the accuracy costs of any "confirmation bias"[9] and makes a link with "motivated reasoning" questionable.

Consideration of systematicity and scope of a putative bias consequently necessitates a clear distinction between the different component processes that go into the formation of a judgment and its subsequent report (whether in an experiment or in the real world). Figure 2.4 distinguishes the three main components of a judgment: evidence accumulation; aggregation, and evaluation of that evidence to form an internal estimate; and report of that estimate. In the context of wishful thinking, biasing effects of outcome utility (the desirability/undesirability of an outcome) can arise at each of these stages (readers familiar with Funder's (1995), realistic accuracy model of person perception will detect the parallels; likewise, motivated reasoning research distinguishes between motivational effects on information accumulation and memory as opposed to effects of processing, see e.g., Kunda, 1990). Figure 2.4 provides examples of studies concerned with biasing effects of outcome desirability on judgment for each of these component processes. For instance, demonstrations that participants' use information about real-world base rate (Dai et al., 2008) or real world "representativeness" (Mandel, 2008) in judging the probability of events exemplify effects of outcome utility on the information available for the judgment: events that are extremely bad or extremely good are less likely in the real world than ones of moderate desirability, so that outcome utility provides information about frequency of occurrence which can be used to supplement judgments where participants are uncertain about their estimates.

[8] As a reminder, the target rule governing triples of numbers in Wason's study was "increasing numbers." A positive test strategy means testing triples that would be instances of the currently preferred rule. This cannot lead to success when the true rule is less general than the current hypothesis (e.g., "increasing by two" vs. "increasing numbers").

[9] Though people may still, and most likely do, do less well than an optimal model by overreliance on positive test strategies even in circumstances where its expectation is lower than that of a negative test strategy (see for some examples, Klayman & Ha, 1989).

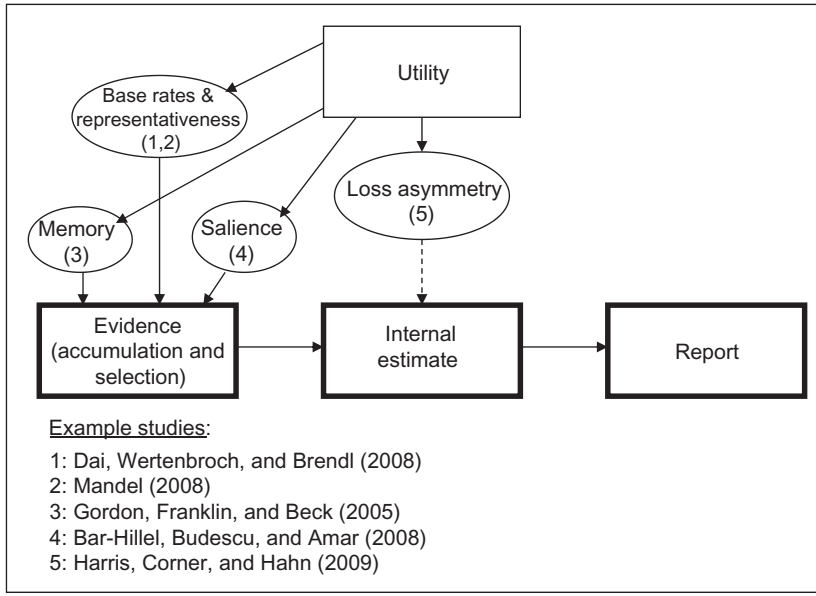

Figure 2.4 Locating indirect effects of utility (outcome desirability/undesirability) in the probability estimation process. Framed boxes indicate the distinct stages of the judgment formation process. Ovals indicate factors influencing those stages via which outcome utility can come to exert an effect on judgment. Numbers indicate experimental studies providing evidence for a biasing influence of that factor. Note that Dai, Wertenbroch, and Brendl (2008), Mandel (2008), and Harris et al. (2009) all find higher estimates for undesirable outcomes (i.e., "pessimism"). *Figure adapted from Harris et al. (2009).*

Confirming our observations about the relative reliability of primary and secondary bias in generating systematic deviations, the different components of the judgment process vary in the extent to which they generally produce "wishful thinking" and several of the studies listed (see Fig. 2.3) have actually found "anti" wishful thinking effects, whereby undesirable events were perceived to be more likely.

Such mixed, seemingly conflicting, findings are, as we have noted repeatedly, a typical feature of research on biases (see e.g., Table 1 in Krueger & Funder, 2004). However, only when research has established that a deviation is systematic has the existence of a bias been confirmed and only then can the nature of that bias be examined. The example of base rate neglect above illustrated how examination of only a selective range of base rates (just low prior probabilities or just high prior probabilities) would have led to directly conflicting "biases." The same applies to other putative biases.

In general, names of biases typically imply a putative scope: "wishful thinking" implies that, across a broad range of circumstances, thinking is "wishful." Likewise, "optimistic bias" (a particular type of wishful thinking, see Sharot, 2012) implies that individuals' assessments of their future are *generally* "optimistic." Researchers have been keen to posit broad scope biases that subsequently do not seem to hold over the full range of contexts implied by their name. This suggests, first and foremost that no such bias exists.

To qualify as optimistically biased for example, participants should demonstrate a tendency to be optimistic across a gamut of judgments or at least across a particular class of judgments such as probability judgments about future life events (e.g., Weinstein, 1980; in keeping with Weinstein's original work we restrict the term "optimistic bias" to judgments about future life events in the remainder). However, while people typically seem optimistic for rare negative events and common positive events, the same measures show pessimism for common negative events and rare common events (Chambers et al., 2003; Kruger & Burrus, 2004). Likewise, for the better-than-average effect (e.g., Dunning, Heath, & Suls, 2004; Svenson, 1981), people typically think that they are better than their peers at easy tasks, but worse than their peers at difficult tasks (Kruger, 1999; Moore, 2007), and the false consensus effect (whereby people overestimate the extent to which others share their opinions, Ross, Greene, & House, 1977) is mirrored by the false uniqueness effect (Frable, 1993; Mullen, Dovidio, Johnson, & Copper, 1992; Suls, Wan, & Sanders, 1988).

One (popular) strategy for responding to such conflicting findings is to retain the generality of the bias but to consider it to manifest only in exactly those situations in which it occurs. Circumstances of seemingly contradictory findings then become "moderators," which require understanding before one can have a full appreciation of the phenomenon under investigation (e.g., Kruger & Savitsky, 2004): in the case of the better-than-average effect therefore that moderator would be the difficulty of the task.

2.1.1 The Pitfalls of Moderators

Moderators can clearly be very influential in theory development, but they must be theoretically derived. *Post hoc* moderation claims ensure the unfalsifiability of science, or at least can make findings pitifully trivial. Consider the result—reported in the Dutch Daily News (August 30th, 2011)—that thinking about meat results in more selfish behavior. As this study has since been retracted—its author Stapel admitting that the data were fabricated—it is likely that this result would not have replicated. After (say) 50 replication attempts, what is the most parsimonious conclusion?

One can either conclude that the effect does not truly exist or posit moderators. After enough replication attempts across multiple situations, the latter strategy will come down to specifying moderators such as "the date, time and experimenter," none of which could be predicted on the basis of an "interesting" underlying theory.

This example is clearly an extreme one. The moderators proposed for the optimism bias and better-than-average effects are clearly more sensible and more general. It is still, however, the case that these moderators must be theoretically justified. If not, "moderators" may prop up a bias that does not exist, thus obscuring the true underlying explanation (much as in the toy example above). In a recent review of the literature, Shepperd, Klein, Waters, and Weinstein (2013) argue for the general ubiquitousness of unrealistic optimism defined as "a favorable difference between the risk estimate a person makes for him- or herself and the risk estimate suggested by a relevant, objective standard. . .Unrealistic optimism also includes comparing oneself to others in an unduly favorable manner," but state that this definition makes "no assumption about why the difference exists. The difference may originate from motivational forces. . .or from cognitive sources, such as. . .egocentric thinking" (Shepperd et al., 2013, p. 396).

However, the question of why the difference exists is critical for understanding what is meant by the term unrealistic optimism especially in the presence of findings that clearly appear inconsistent with certain accounts. The finding that rare negative events invoke comparative optimism, while common negative events invoke comparative pessimism seems entirely inconsistent with a motivational account. If people are motivated to see their futures as "rosy," why should this not be the case for common negative events (or rare positive events) (Chambers, Windschitl, & Suls, 2003; Kruger & Burrus, 2004)? One can say that comparative optimism is moderated by the interaction of event rarity and valence, such that for half the space of possible events pessimism is in fact observed, but would one really want to call this "unrealistic optimism" or an "optimistic bias"? Rather, it seems that a more appropriate explanation is that people focus overly on the self when making comparative judgments (e.g., Chambers et al., 2003; Kruger & Burrus, 2004; see Harris & Hahn, 2009 for an alternative account which can likewise predict this complete pattern of data)—a process that simply has the by-product of optimism under certain situations. It might be that such overfocus on the self gives rise to bias, but through a correct understanding of it one can better predict its implications. Likewise, one is in a better position to judge the potential costs of it.

In summary, when bias is understood in a statistical sense as a property of an expectation, demonstration of deviation across a range of values is essential to establishing the existence of a bias in the first place, let alone understanding its nature. Conflicting findings across a range of values (e.g., rare vs. common events in the case of optimism) suggest an initial misconception of the bias, and any search for moderators must take care to avoid perpetuating that misconception by—unjustifiedly—splitting up into distinct circumstances one common underlying phenomenon (i.e., one bias) which has different effects in different circumstances (for other examples, see on the better-than-average/worse-than-average effect, see e.g., Benoit & Dubra, 2011; Galesic, Olsson, & Rieskamp, 2012; Kruger, 1999; Kruger, Windschitl, Burrus, Fessel, & Chambers, 2008; Moore & Healy, 2008; Moore & Small, 2007; Roy, Liersch, & Broomell, 2013; on the false uniqueness/false consensus effect see Galesic, Olsson, & Rieskamp, 2013; more generally, see also, Hilbert, 2012).

3. MEASURING BIAS: THE IMPORTANCE OF OPTIMAL MODELS

Having stressed the importance of viewing bias as an expected deviation from accuracy, and the attendant need to examine the performance of inferential procedures across a range, we next highlight the difficulties of establishing a deviation from accuracy in the first place.

In some cases, the true value is known and can be compared directly with a participant's estimate, but as we saw earlier, in social psychological studies the true value is typically not known and researchers resort to other, less direct ways of ascertaining bias. Here, comparison with normative models that specify how to perform the task "optimally" is essential.

Given the unease with normative models in social psychology, we not only provide some examples of where common-sense intuition may be misled, but also outline why the normative standard—in this case Bayesian probability—is appropriate to the task.

Our first example carries on with the "wishful thinking" theme by examining unrealistic optimism about future life events.

Here, people's actual risk is not known directly to the participant or the experimenter, rather it must be inferred from available evidence. "Rationality" in such circumstances is thus necessarily about inferential procedures, not about whether particular answers derived by such a procedure are right or wrong as such—simply because it is in the nature of induction as an ampliative inference that goes beyond the data given—that sometimes even one's "best guess" will be wrong (see e.g., Baron, 2008).

3.1. Bayesian Belief Revision

The common standard for such inductive inference is Bayesian belief revision. Within the Bayesian framework, probabilities are conceptualized as subjective degrees of belief, rather than as objective frequencies existing in the external environment. Bayesian inference is thus concerned with the consistency and coherence of those probabilities. Most importantly in the present context, it can be shown that "being Bayesian" (that is adherence to the axioms of the probability calculus and the use of Bayes' rule in incorporating evidence for the revision of ones beliefs) has lawful connections with accuracy. Accuracy-based justifications of Bayesianism invoke scoring rules as are used to measure the accuracy of probabilistic forecasts (e.g., in meteorology). As shown by de Finetti (1974), given a scoring rule by which a person incurs a penalty of $(1-p)^2$ if an event is found to be true and p^2 if an event is found to be false (where p denotes a numerical value previously assigned by the person to the likelihood of the event in question), a person will *necessarily* incur a larger penalty if their likelihood estimates do not obey the probability axioms. Lindley (1982, 1994) argues that if other scoring rules are used then either people should provide responses that are, in reality, only transformations of probability (e.g., odds), or people should only estimate 0 or 1 (demonstrating the inadequacy of such a scoring rule). Hence, "all sensible rules lead back, via a possible transformation, to probability. Probability is inevitable" (Lindley, 1994, p. 6; see also, e.g., Cox, 1946; Horvitz, Heckerman, & Langlotz, 1986; Snow, 1998).

Furthermore, Rosenkrantz (1992) shows that updating by Bayes' rule maximizes the expected score after sampling; in other words, other updating rules will be less efficient in the sense that they will require larger samples, on average, to be as accurate. Leitgeb and Pettigrew (2010b) demonstrate that for a suitable measure of accuracy (justified in Leitgeb & Pettigrew, 2010a; but see Levinstein, 2012), Bayesianism follows from the simple premise that an agent ought to approximate the truth, and hence seek to minimize inaccuracy.

Being Bayesian thus provides a route for realistic beliefs about future life events as studied within the unrealistic optimism literature. How then does a Bayesian form their beliefs about whether or not they will at some point in their lives experience a negative life event? Bayes' rule provides normative guidance on how beliefs should be updated upon receipt of new information:

$$P(h|e) = \frac{P(h)P(e|h)}{P(e)} \tag{2.2}$$

Equation (2.2) mandates that one evaluate the likelihood of a hypothesis, h, in light of some evidence, e, by multiplicatively combining one's prior degree of belief in that hypothesis, $P(h)$ (i.e., one's degree of belief before receiving the present evidence), with the likelihood of obtaining that evidence in the case that the hypothesis is true, $P(e|h)$, and then normalize by dividing by the likelihood of obtaining that piece of evidence regardless of the truth or falsity of the evidence, $P(e)$.

The base rate of an event (let us use a disease as an example) in a population can be defined as a frequentist percentage—for example, the number of people per 100 who will contract the disease. This is the most appropriate information to use as one's prior degree of belief that the disease will be contracted, $P(h)$. Consequently, if no individuals within the population have any information with which to differentiate their chance of experiencing a disease from the average person's, all individuals should state that their chance of experiencing the disease is equal to the base rate. It is easy to see in this instance that the population's probability judgments are coherent and well calibrated: each individual states their chance of contracting the disease as being the base rate; thus, the average-risk rating of the population is the base rate, which in turn is the same as the true proportion of people who will actually experience the disease.

However, this relationship will also hold in instances in which different individuals have information to differentiate their own chance from the average person's. Consider a simple situation in which a disease is known to be predictable, in part, by family history (as is the case for many serious illnesses, e.g., Walter & Emery, 2006). We further assume that everyone in the population knows whether or not they have a family history of the disease and that this is the only information that is known about the etiology of the disease. Those with a family history will estimate their own risk using Eq. (2.2), while those without will be estimating the likelihood that they will contract the disease knowing they have no family history of it, $P(h|\neg e)$ (Eq. 2.3):

$$P(h|\neg e) = \frac{P(h)P(\neg e|h)}{P(\neg e)} \tag{2.3}$$

The average of people's estimates of self-risk is obtained by adding together those estimates and dividing by the number of individuals. This is functionally equivalent to multiplying each of the two risk estimates (Eqs. 2.2 and 2.3) by the proportion of people expressing them and summing

across the two distinct (average) estimates; the result will once again equal the base rate (Harris & Hahn, 2011). Thus, the average risk estimate of a population of Bayesians who use the base rate to estimate their prior degree of belief will result in average risk estimates that are calibrated (at the population level) to the base rate statistic.

This guaranteed population level calibration demonstrates why Bayesian belief revision is normatively appropriate in this context and it confirms researchers' intuition that optimism can be assessed at the group level even in the absence of knowledge about a given individual's personal likelihood of experiencing an event.

3.2. Divergence of Normative Predictions and Experimenter Intuition

3.2.1 Unrealistic Comparative Optimism

At a group level, average estimates of personal risk should equal average estimates of the average person's risk, or base rate, $P(h)$, whether we know the individual's risk and the true base rate or not. Consequently, if the population's estimate of self-risk is *lower* than their estimate of the base rate then their estimates can be said to be biased in an optimistic direction—the classic phenomenon of unrealistic comparative optimism (e.g., Weinstein, 1980).

While the basic logic of this is sound, its application in standard experimental methods is not. Harris and Hahn (2011) demonstrated that optimal Bayesian agents completing standard questionnaires for studying optimism (following Weinstein's classic, 1980, method) would show patterns taken to indicate unrealistic optimism, even though they are entirely rational and in no way optimistic. This stems from three independent sources whose effects will combine and amplify one another in any given study: (1) scale attenuation, because participants provide their responses on a limited, noncontinuous, scale (e.g., −3 to +3, see e.g., Covey & Davies, 2004; Klar, Medding, & Sarel, 1996; Price, Pentecost, & Voth, 2002; Weinstein, 1982, 1984, 1987; Weinstein & Klein, 1995) which does not allow faithful conversion of underlying probability estimates; (2) minority undersampling, arising from the fact that estimates are elicited from only a sample taken from the population so that calibration at the group level is no longer guaranteed; and (3) base rate regression. The base rate regression mechanism derives from the fact that people's actual estimates of probabilities are likely to be regressive; that is, due to random error small probabilities will, on average, be overestimated and large probabilities will be underestimated (because of the bounded nature of the probability scale). This will

bias population results because base rates assumed by agents in the calculation of their individual risk will no longer match the impact of diagnostic evidence. Bayes' rule no longer guarantees that the appropriate incorporation of diagnostic evidence results in population level calibration if the base rate estimates going into the agents' estimates are regressive. Specifically, their mean will no longer equal the average individual risk, because that is based not just on base rate estimates but also on the diagnostic evidence each individual has received, and this diagnostic evidence is dispensed "by the world," as it were.[10] It is thus governed by the true base rate, not by its regressive subjective estimate. Put simply, the number of people who will have a family history of a disease as opposed to the number who will not depends on the "true" distribution of the disease, not one's beliefs about it. Harris and Hahn (2011) demonstrate that the resultant mismatch will yield comparative optimism for rare negative events (i.e., mean estimates of self-risk are lower than the perceived "average person's risk") and absolute optimism (i.e., the mean estimated self-risk is higher than the true population mean), as is indeed observed in the empirical literature (Chambers et al., 2003; Kruger & Burrus, 2004; Moore & Small, 2008).

The same underlying issue of a clash between the diagnostic information dispensed "by the world" and the base rates that are assumed (by participants or experimenters) plagues research that has sought to provide evidence of unrealistic optimism by studying belief updating directly, as opposed to examining its results.

3.2.2 Optimistic Belief Updating

Lench and colleagues' "automatic optimism" (Lench, 2009; Lench & Bench, 2012; Lench & Ditto, 2008) provides a motivational account of optimism bias based on an "optimism heuristic." As a default reaction, people "simply decide that events that elicit positive affective reactions are likely to occur and events that elicit negative affective reactions are unlikely to occur" (Lench & Bench, 2012, p. 351), when they experience approach or avoidance reactions to future events. In support, Lench and Ditto (2008) provided participants with matched positive and negative future life events, for which participants were given equal base rates. Participants then

[10] Shepperd et al. (2013) appear to misunderstand this base rate regression mechanism, confusing it with accounts of other self-enhancement phenomena in terms of differentially regressive estimates for self and other (e.g., Moore & Small, 2008). Differential regression concerns the relationship between two estimates, one of which is more regressive than the other. Other than that "regressiveness" of ratings are involved in both differential regression and the base rate regression mechanism, they have nothing to do with one another and are not only conceptually but also empirically distinct.

rated their own chance of experiencing that event. In a direct comparison, the mean estimates for negative events were lower than for the matched positive events, suggesting optimism.

The problem lies in the fact that Lench and Ditto's (2008) used negation to generate corresponding negative ("will get cancer") and positive ("will not get cancer") events. However, complementary events can be equiprobable only if their base rate is exactly 50%. This is not the case for events such as "getting cancer," "owning one's own home," "at some point being unemployed," or "developing asthma" as used by Lench and Ditto (2008). That participants are told equivalent base rates does not make things equal, because the distribution of participants' individual diagnostic knowledge will be governed by the true base rate (i.e., "the way the world is"), precisely because that knowledge is diagnostic.

To illustrate, cancer has a life-time prevalence of about 40%, so most people will genuinely not get cancer. By the same token, more people will possess diagnostic knowledge indicating lower risk than there will be people with knowledge indicating greater risk. This means that averages across estimates of individual risk will deviate from the experimenter provided base rate even if participants fully believe that base rate and incorporate it into their own prediction. Moreover, because the negative event in question is rare, it will necessarily deviate in the direction of seeming "optimism," even if people's estimates are otherwise fully rational and Bayesian.

The same issue plagues Sharot and colleagues demonstrations of seemingly optimistic belief updating and its moderators (Chowdhury, Sharot, Wolfe, Düzel, & Dolan, 2013; Korn, Sharot, Walter, Heekeren, & Dolan, 2014; Moutsiana et al., 2013; Sharot, Guitart-Masip, Korn, Chowdhury, & Dolan, 2012; Sharot, Kanai, et al., 2012; Sharot, Korn, & Dolan, 2011). In these studies, participants seemingly display motivated reasoning by virtue of the fact that they are selective in their use of desirable and undesirable information, revising their beliefs to a greater extent in response to "good news" than to "bad."

Participants are required to estimate their chance of experiencing a series of negative events. As in Lench and Ditto's (2008) study, they are then provided with the base rate statistic for each event, in this case, a genuine estimate of the true base rate.[11] Participants then provide a second estimate of their personal risk. Rather than comparing the means of those estimates with

[11] This "true base rate" is a base rate sourced from real-world statistics about people from the same sociocultural environment as the participants, although given that sources for these statistics include the Office for National Statistics, it is unlikely that University College London research participants (e.g., Sharot et al., 2011) constitute a representative sample of the general population from which that estimate was devised.

the base rate, Sharot and colleagues examine the amount of belief change from first to second estimate. The typical finding is a seemingly optimistic asymmetry in belief updating. Specifically, when the chance of experiencing a negative event is higher than participants initially thought (undesirable information), they revise their personal risk estimates less than when it is lower (desirable information).

This result seems a particularly strong form of motivated reasoning, since it is difficult to envisage how participants could maintain any "illusion of objectivity" given they receive *only* the base rate (e.g., Kunda, 1990). Motivated reasoning research has typically come to the conclusion that, psychologically, people do not seem at liberty to distort their beliefs however they desire; rather, their motivational distortions must have at least some basis in the evidence allowing them to maintain the "illusion of objectivity" by selectively focusing on particular aspects in order to reach a desired conclusion. Recent demonstrations that participants updating are asymmetrically optimistic in their judgments about their attractiveness (Eil & Rao, 2011) and their intelligence (Eil & Rao, 2011; Mobius, Niederle, Niehaus, & Rosenblat, 2011), seem more in line with this typical view; attractiveness is a highly subjective attribute, and intelligence is a multidimensional construct, of which different intelligence tests typically measure different aspects, giving participants ample room for selective focus that seems unavailable in Sharot et al.'s paradigm.

However, Sharot et al.'s paradigm also differs methodologically from these recent studies of belief revision about intelligence or attractiveness. Whereas Eil and Rao (2011) and Mobius et al. (2011) compare actual updating with explicitly calculated Bayesian prescriptions, Sharot et al. simply measure belief change. Unfortunately, this is a good example in which intuition and normative standards clash, and again, the distribution of events in the world (and with them diagnostic information) relative to perceived base rates is key to where intuition goes wrong.

Given that the only "new information" participants receive in Sharot et al.'s paradigm is the "true" base rate, this is also the only quantity they should modify in calculating their own best estimates. Rational agents updating their beliefs in accordance with Bayes' rule should replace their own base rate estimate with the true base rate and recalculate their estimate of individual risk (Eq. 2.3). If their initial base rate estimate was lower than the true base rate, then their new estimate of self-risk will be higher; if their initial self-estimate was based on an overestimate of the base rate, then their new estimate of self-risk will be lower; and finally, if the two base rates match, their estimate remains unchanged.

It seems plausible that given equal amounts of deviation from the base rate for both over- and underestimates, the change scores (that is the absolute value of the first self-estimate–second self-estimate) should be equal between those receiving "good news" (a lower base rate than they had assumed) and those receiving "bad news" (a higher base rate than assumed). However, this is *not* the case. Equal (average) error in base rate estimate does not translate directly into equal average error in self-estimates (and hence "change" on correction), because—once again—the distribution of diagnostic information follows the true base rate, not the perceived base rates. Consequently, even for unbiased error about the base rate (i.e., mean = 0) average change for entirely rational agents can differ between base rate over- and underestimators (see Shah, Harris, Bird, Catmur, & Hahn, 2013, Appendix B for demonstration). "Equal change in self-estimate" across good and bad news about the base rate is thus *not* normatively mandated, and can fail in entirely rational, Bayesian, agents.

Sharot and colleagues' actual studies further compound this methodological error by failing to ascertain participants' initial base rate estimates in order to determine whether they are, in fact, receiving good or bad news about the base rate. Rather they identify "good news" and "bad news" relative to the participants' estimate of *individual* risk, but that estimate, of course, rightfully also contains individual, diagnostic information. Rational agents in possession of diagnostic information indicating lower risk (e.g., those who do not have a family history of cancer) legitimately consider themselves to be at less risk than the "average person" (see also Weinstein & Klein, 1996), and the only "news" they have received concerns the base rate. "Good" and "bad" news must consequently be defined relative to that base rate, not relative to perceived self-risk.

Analytic examples and simulations demonstrate that classification of entirely rational Bayesian agents on the basis of the relationship between self-estimate and base rate yields considerable misclassification, and is sufficient to generate data indicative of "optimistic updating" even for a population of simulated rational Bayesian agents (Harris, Shah, Catmur, Bird, & Hahn, 2013).

It is worth emphasising that nothing in the preceding analyses requires that participants actually be Bayesian. Rather the point is that an experimental measure that yields "optimism" with rational agents who are, by design, rational and not optimistic, can provide neither evidence of optimism nor of irrationality.

In summary, intuitions about accuracy can go badly wrong, making consideration of normative, optimal models essential. In reference to debates and concerns about normative status, it is important to acknowledge that such

debates exist and that the theoretical understanding of norms of rationality is incomplete and still developing. This makes it important to articulate in a given context why something is considered a norm and what relationship its justification bears to the problem at hand (see also Corner & Hahn, 2013).

One issue that has been subject to debate is the interpretation of probability (and hence risk) itself. On a Bayesian, subjectivist interpretation, probabilities reflect subjective degrees of belief, whereas on a frequentist, objectivist interpretation of probability, they reflect proportions (see e.g., von Mises, 1957/1981). The two are not mutually exclusive to the extent that Bayesians also wish to adopt "objective probabilities" (such as the proportion of balls of a given color in an urn from which random draws are taken) as subjective degrees of belief, where such probabilities are available (the so-called Principal principle, Lewis, 1986; Meacham, 2010), and this is what we have assumed in our examples of populations and adverse events such as diseases. However, where Bayesians and frequentists differ is in probabilities for unique events, such as whether or not I will contract a particular disease at some point in my life. For frequentists, there is no sample from which proportions could be inferred, so questions about the probability of singular future events are meaningless (an issue that has figured prominently in critiques of the Heuristics and Biases program, see e.g., Gigerenzer et al., 1989; Cosmides & Tooby, 1996). They can be understood only as questions about individuals as members of a reference class (e.g., "women," "women who smoke," "women who are overweight and smoke," etc.). From this perspective, questions about self- and average risks in unrealistic optimism studies involve different reference classes, where self-risk may be taken from a more specific reference class ("men with no family history of the disease who exercise regularly") than the average person (i.e., the target population). Again, ratings of self-risk may legitimately be lower than average-risk ratings. Moreover, due to what is known as the "reference class problem" there is no unique answer for self-risk: any given individual will belong to many different references classes (i.e., "woman who are overweight and smoke" are also "woman who smoke" and "woman," see e.g., Cosmides & Tooby, 1996). As a consequence, there is also no reason why self-ratings should average out and equal ratings of average risk,[12] so, from a purely frequentist perspective, optimism research, whether it be comparative ratings or belief updating, is a nonstarter. In short,

[12] Specifically, average self-risk would average out to the population mean only if all reference classes used by raters formed a partition, that is, were mutually exclusive and jointly exhaustive of the population as a whole. There is no way to ensure in standard test that they ever would.

competing strands of the normative debate concur in leading to a negative assessment of these methods.

In conclusion, we could not agree more with "norm sceptics" that simply claiming that something is normative is not enough. However, as the present examples hopefully illustrate, the fact that there can be legitimate debate about normativity does not preclude that, in a specific context, the issues may be clear enough, and that it would be detrimental to ignore putatively normative considerations: an incomplete map of the terrain is likely to still be better than no map at all.

3.3. Bayes and Experimental Demonstrations of Motivated Reasoning

Our critical evaluation of biased focused research in the context of motivated reasoning thus far has focused on only a subset of what is a vast and sprawling literature. Wishful thinking in the context of judgments of probability or risk constitutes only a small portion of the relevant research literature; at the same time, however, it is a part which, in principle, affords rigorous research with respect to bias. Systematicity of deviation from expected values can, and has been, evaluated in that researchers have examined many kinds of outcomes over a broad range of probabilities where criteria for accuracy exist at least in principle.

It is our contention that the majority of the remaining research on motivated reasoning has not done enough to establish "bias" in the sense of systematic deviation from accuracy—let alone establish that participants' reasoning is irrational or flawed. This is obviously a contentious statement, but it follows rather directly from the fact that most studies of motivated reasoning that fall outside the paradigms already discussed rely on the impact of what are perceived to be normatively irrelevant experimental manipulations on participants' beliefs as their methodology. Not only does that bring with it evaluations of performance that lack systematic variations of context, but it also means that the purported impact of participant sensitivity to such manipulations stands and falls with experimenter intuitions, which are typically not given any robust justification. Consideration of fundamental aspects of Bayesian belief revision, in many cases, suggests that these experimenter intuitions are hard to defend.

In this context, it is qualitative properties of Bayesian belief revision that are relevant, simply because most of the experimental studies show only that responses are "different" across the conditions. Matching broad qualitative prescriptions of Bayes' rule is obviously a good deal easier than matching

quantitatively the precise degree to which beliefs should change. Thus, our analysis in this section leaves open the possibility that people may fail more exacting quantitative tests. Indeed, this is to be expected in light of the detailed findings of less than optimal sensitivity to probabilistically relevant variables ("conservatism") within the 1960s tradition of bookbags and pokerchips. We are thus by no means out to proclaim complete rationality of participants; rather, the purpose is to point out that, if responding is in keeping with broad qualitative trends, then establishing bias must by necessity go down more specific and detailed routes. At the same time, however, qualitative sensitivity to "the right factors" will serve to bound participants' inaccuracy in practice.

Finally, in keeping with earlier emphasis on the need to be explicit about one's normative justifications, it seems relevant to point out (in addition to the reasons given for adherence to the Bayesian calculus so far) that the Bayesian framework has come to take a central role in current epistemology and philosophy of science as a standard for rationality (see e.g., Bovens & Hartmann, 2004; Howson & Urbach, 1996). This, in and of itself, seems enough to support the perception that patterns of inference that are in qualitative agreement with Bayesian prescriptions are not obviously *irrational* whatever experimenters may have assumed!

With these introductory words in place, what can be concluded about bias on the basis of the motivated reasoning literature?

Clearly, a systematic review of that literature is beyond the scope of this chapter. So our examples will necessarily be selective (though hopefully with broader implications). In this case, it seems appropriate to take as our point of departure key reviews of that literature; thus, whatever general picture is drawn in those, it will at least not be driven by our own perspective. The standard point of departure here is Kunda's classic (1990) review.

Kunda's review is set in the historical context of long-standing debate between cognitive and motivational explanations of findings, in particular in the context of attribution that seemed to indicate motives affecting reasoning in such a way as "to allow people to believe what they want to believe because they want to believe it" (Kunda, p. 480). Critiques of motivated explanations of such findings maintained that early findings could be understood entirely in nonmotivational, cognitive terms (e.g., Dawes, 1976; Miller & Ross, 1975; Tetlock & Levi, 1982). Kunda's own conclusion, a decade later, was that whereas these earlier critiques rejected the case for motivational forces on parsimony grounds (as findings were explicable in cognitive terms alone), the situation had now reversed in that "a single

motivational process for which unequivocal independent evidence now exists may be used to account for a wide diversity of phenomena (p. 493)," many of which could not be accounted for in nonmotivational, cognitive terms, or would require ad hoc assumptions without independent support.

Our focus is on accuracy and bias (in the sense of systematic deviations from accuracy); consequently, the distinction between cognitive and motivational factors is of interest, here, only to the extent that it might reliably be associated with differential outcomes with regard to accuracy.

Kunda's main conclusions are that people's inferential processes are subject to two motivational influences: (1) a motivation to be accurate and (2) a motivation to reach a desired conclusion. Moreover, on the available evidence, even directionally motivated reasoning does not constitute a carte blanche to believe whatever one desires; the desired conclusion is only drawn if it can be supported by evidence—indeed, if that evidence could "persuade a dispassionate observer" (Kunda, 1990, pp. 482–483). Kunda provides only very limited evidence of judgmental distortions, and what evidence is listed is quite weak (e.g., purported evidence of direct biasing influences of desirability on probability judgement, which subsequent research on "wishful thinking" as discussed above has discredited). Rather, in her view, the key mechanism that past research points to is accumulation and selection of evidence that is biased in such a way that the resulting inferential process might lead to an outcome that seems biased from an external perspective (i.e., viewed in terms of correspondence), but which is subjectively rational given the evidence considered at that time (i.e., viewed in terms of coherence, at least in terms of the selected evidential base). In fact, she acknowledges that present evidence is entirely compatible with the idea that the impact of motivation on reasoning is exhausted by the setting of a directional query or hypothesis (e.g., "Am I healthy?" as opposed to "Am I ill?") without further effect on the processes through which these questions are answered.

We are thus led back to the issue of the extent to which biases at the information accumulation stage may reliably support a *systematic* deviation from accuracy, as opposed to occasional error.

However, it also remains less than clear to what extent biases at these stages, in fact, exist. With regard to evidence accumulation, Hart et al. (2009) conducted an extensive meta-analysis of 67 experimental studies examining "selective exposure," that is, the extent to which people choose to examine information they have been told will be congenial to a prior

attitude or belief they hold, as opposed to information that runs counter to it. As discussed above, this perfect (or near perfect) correlation between an information source and its content is not typically found in the real world. In the real world, we do not have this much control over the content of the information we receive, but such paradigms may nevertheless be indicative of selectivity. The 67 studies provide both evidence of a congeniality bias and evidence for its opposite, an anticongeniality bias. This not only raises the now familiar question about the existence of a congeniality bias, but it also, in itself, lessens the probability of systematic impact on the accuracy of beliefs.

Hart et al. (2009) do go on to examine both the shape of the distribution of effect sizes and to calculate an average effect size (which they find to be positive and indicative of participants, on average, being almost twice as likely to select congenial over uncongenial information). However, with regard to bias in the sense of expected deviation from a true value, the analysis makes no distinctions between studies examining beliefs about facts, for which truth, and hence accuracy as correspondence with the truth is well defined, and attitudes, for which no objective standards may be available. This makes sense in the wider context of motivated reasoning research which has not distinguished between beliefs and attitudes, but it is essential to the question of rationality and quantifying deviations from true values for establishing bias.[13] Moreover, it is important not just at the point of trying to calculate such values, but also at the point of examining behavior: it cannot be assumed that people's information acquisition and evaluation strategies are the same whether the target concerns a fact or the degree to which someone or something is "liked," and there is much reason, both normative and empirical, to assume that they are not. Consequently, including in the calculation of effect sizes studies for which a correct answer may not be defined clouds the extent to which "congeniality bias" exists in a form that could negatively affect accuracy even in principle.

The same applies to studies of congeniality bias in memory in the context of attitudes. Here, too, the meta-analysis by Eagly et al. (1999) reveals considerable inconsistency in the literature, with an overall meta-analytic effect size of the congeniality effect in memory of zero (with 40% of studies showing the opposite bias, an uncongeniality effect). This latter result led Eagly

[13] To be clear, the congeniality question in the context of valuation is of course of central importance to "bias" in the sense of impartiality or fairness, but the focus of the present chapter is on whether people are rational, not on whether people are nice.

et al. to propose that people engage in "active" defense strategies. That is, attitude inconsistent information is attended to at least as much as attitude inconsistent information, and processed more deeply to enable counter-arguments (support for which was obtained in Eagly, Kulesa, Brannon, Shaw, & Hutson-Comeaux, 2000, Eagly, Kulesa, Chen, & Chaiken, 2001; see also, Ditto, Scepansky, Munro, Apanovitch, & Lockhart, 1998; Edwards & Smith, 1996). More recently, Waldum and Sahakyan (2012) found that both directed memory and directed forgetting were enhanced for incongruent versus congruent political statements, which seemed based on more episodic contextual information being encoded in the memory trace for incongruent versus congruent information.

In either case, whether or not the true effect size for congeniality bias for beliefs in exposure or memory is zero or not, the fact that it is so strongly subject to "moderating factors," again, weakens the extent to which it could have *systematic directional* effects on our beliefs, as opposed to promoting occasional error.

A final question, however, remains, and that is the question of what, from a normative perspective, would actually promote accuracy goals and hence what should count as criteria for "defensive" or "accuracy" orientation, whether in the selection or the subsequent judgmental evaluation of evidence. The standard view in the literature on motivated reasoning is this:

> *Accuracy motivation should promote tendencies to process information in an objective, open-minded fashion that fosters un-covering the truth*
> ***(Chaiken et al., 1989; Kunda, 1990) (quoted from Hart et al., 2009, p. 558).***

The assumption, here, is that it is some kind of even-handedness or objectivity that is critical to accuracy in our inferences. The same intuition is present in the "neutral evidence principle." Mixed evidence in the context of biased assimilation paradigms, such as in Lord et al.'s (1979) study, should not change our beliefs, because positive and negative evidence should balance each other out; that is, regardless of our prior beliefs, the diagnostic impact of a piece of evidence should be the same.

Following the discussion of bias in statistics above, in particular the comparisons between a Bayesian and a classical estimator of proportion, it should already be apparent that this is too simplistic. The impact of a piece of evidence is not constant across the range of priors, and Bayesian inference has its relationship with accuracy not just in spite of the fact that judgment is influenced by priors, but also because of it. Where information is received sequentially, that is bit by bit, as is typical in the world, priors summarize past evidence.

One may argue, however, that even though the actual effect on our beliefs of a piece of evidence may, from a Bayesian perspective, vary, its diagnosticity, as measured by the LHR, should at least stay the same. That is, wherever a piece of positive information takes us to, an equally diagnostic piece of negative information should take us back to where we started—in keeping with the neutral evidence principle.

However, even this may—and has been—disputed from a normative perspective. In particular, it is questionable whenever source reliability comes into play. Much of our knowledge comes from the testimony of others. We have, for example, typically not observed the outcome of scientific or medical tests directly, but rather have access to them only through reports. Moreover, this is exactly the situation in which experimental participants in motivated reasoning experiments find themselves.

After a long history of neglect, philosophical interest has recently turned to testimony (e.g., Coady, 1992) and a crucial aspect of testimony is trust. Unless, we believe a source to be perfectly reliable—a condition unlikely to be met by even the most well-intentioned informants—the impact of testimonial evidence should be somewhat less than had we observed the evidence directly (see e.g., Hahn, Harris, & Corner, 2009; Hahn, Oaksford, & Harris, 2012; Schum, 1981, 1994). From a Bayesian, epistemological perspective, source, and evidence characteristics combine to determine the overall diagnostic value of the evidence. Furthermore, the content of the testimony itself may provide one indicator (and in many contexts our only indicator) of the source's reliability. Recent work in epistemology has thus endorsed the position that message content should impact our beliefs about the source (see e.g., Bovens & Hartmann, 2004; Olsson, 2013).

Evidence that is surprising (i.e., conflicts with our prior beliefs) may lower our degree of belief, but it will also lower our degree of trust in the reliability of the source. Although this question has received little attention in psychology, there is some recent evidence to suggest that people naturally draw inferences about the reliability of the source from the degree to which message content is expected (Jarvstad & Hahn, 2011).

This conflicts directly with Baron's neutral evidence principle: once there is no normative requirement for people with opposing views on the content of the message perceive its source as equally reliable, there is also no longer a requirement that they perceive the overall diagnostic value to be the same. What, to an experimenter, may seem equally strong evidence for and against need not be for other observers once source reliability is taken into account.

"Biased assimilation" is indeed a consequence of this, and on occasion, this can lead us to go badly wrong: we end up believing a falsehood while, wrongly, viewing sources who conflict with our opinion as unreliable. However, as we have stressed throughout this chapter, this must distinguished from whether or not this is detrimental to our beliefs *on average*. The fact that an inductive procedure fails occasionally does not mean it is undesirable in general.

The question of the global impact of such sensitivity to source reliability can be examined through simulation. Olsson (2013) simulates a population of Bayesian agents who receive both information "from the world" and from the testimony of others, updating their beliefs about the content of the report and about other's reliability as a function of that content. In the simulation, a proportion of the population ends up believing the wrong thing and distrusting all non-like-minded agents. The majority, however, converge on the truth. The simulation thus shows both "belief polarization" and that such polarization need not undermine our overall accuracy goals (see also, Olsson & Vallinder, 2013). There is much more to be said here than present space permits. Intuitively, the reader may consider that in Lord et al. (1979) study, biased assimilation means that some participants are now "more wrong" than before (depending on which view is actually correct), but those with the opposing view will have moved their beliefs in the direction of "the truth." On average, accuracy may thus readily increase. In summary, it is neither clear that the "neutral evidence principle" is indeed a normative principle, nor that it serves our accuracy goals to be "objective" in the sense that Hart et al. (2009) suggest.

What holds for the judgmental impact of our beliefs, however, also carries through to information selection, and hence, "exposure paradigms." Choosing what information to sample is a *decision*, and thus, normatively subject to expected value. Where the impact of evidence, once obtained, differs, so does its expected value. Source reliability considerations thus affect both "biased" evaluation and selective exposure, and, it would seem that what counts in both contexts as "defensive" as opposed to "accuracy seeking" needs reevaluation.

In summary, consideration of qualitative aspects of Bayesian belief revision indicates that present evidence for motivated reasoning is considerably less good than presumed.

4. CONCLUSIONS

Our "tour" of bias research has, in some ways, come full circle. Source considerations were mentioned as one possible explanation of the "inertia

effect" within 1960s conservatism studies (Peterson & DuCharme, 1967), in that participants may "disbelieve" later evidence (see also, Slovic & Lichtenstein, 1971). Source reliability also provides a potential factor in conservatism more generally (see Corner, Harris, & Hahn, 2010).

It should be clear from the preceding evidence indicating the importance of normative models in studying bias that we think the 1960s studies within the bookbag and pokerchip tradition have much to recommend them. Last but not least, their quantitative nature allows simultaneous assessment both of how bad and how good human judgment is (cf. Funder, 1995; Krueger & Funder, 2004) and affords insight into bias in the all-important sense of systematic deviation from accuracy, alongside assessment of its costs.

The bookbag and pokerchip paradigm has been criticized both on grounds that it is confusing for participants and that it is typically quite artificial and unnatural (e.g., Manz, 1970; Slovic & Lichtenstein, 1971). However, the artificiality does, in fact, make it informative for the study of motivated reasoning. While phenomena such as undersampling and "inertia" (Pitz et al., 1967) are typically cited as evidence in favor of motivated cognition (see e.g., Baron, 2008; Nickerson, 1998), it seems in many ways hard to imagine testing beliefs in which participants could be *less* invested in in any genuine sense, than whether the experimenter-selected bag on this trial contains predominantly red or blue chips. If anything, we thus take the parallels to motivated reasoning phenomena observed in these studies to be evidence *against* motivational accounts. Or to put it differently, if attachments to hypotheses (and with that directional questions) are so readily formed, it, once again, becomes hard to see how motivated cognition could exert any systematic effects on the accuracy of our beliefs. It should also be stressed that it is entirely possible to conduct quantitative studies of belief revision with more naturalistic materials (see e.g., Harris & Hahn, 2009; Harris, Hsu, & Madsen, 2012). Such research, we think, will be necessary, because although some cognitive and social psychologists have recognized and stressed the need to examine global accuracy when studying bias, the majority of this research has not.

The main thing to take away from our critical survey of research on bias is that with respect to the question of human rationality, an interesting notion of bias is established only once it has been shown that there is systematic deviation, that is deviation *on average* across a broad range of instances, and that deviation comes at an accuracy cost, in that there exist actual procedures that could do better. Common-sense intuition, time and again, provides an unreliable guide to when that might be.

Consequently, the rather surprising conclusion from a century of research purporting to show humans as poor at judgment and decision-making, prone to motivational distortions, and inherently irrational is that it is far from clear to what extent human cognition exhibits systematic bias that comes with a genuine accuracy cost.

ACKNOWLEDGMENT

The first author was supported by the Swedish Research Council's Hesselgren Professorship.

REFERENCES

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, *82*, 261–277.

Babad, E., & Katz, Y. (1991). Wishful thinking—Against all odds. *Journal of Applied Social Psychology*, *21*, 1921–1938.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211–233.

Bar-Hillel, M., & Budescu, D. (1995). The elusive wishful thinking effect. *Thinking and Reasoning*, *1*, 71–103.

Bar-Hillel, M., Budescu, D. V., & Amar, M. (2008). Predicting World Cup results: Do goals seem more likely when they pay off? *Psychonomic Bulletin & Review*, *15*, 278–283.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning*, 7, 221–235.

Baron, J. (2008). *Thinking and deciding*. Cambridge: Cambridge University Press.

Benoit, J.-P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica*, *79*, 1591–1625.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Bolstad, W. M. (2004). *Introduction to Bayesian statistics*. Hoboken, NJ: Wiley.

Bovens, L., & Hartmann, S. (2004). *Bayesian epistemology*. Oxford: Oxford University Press.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*, 409–432.

Chambers, J. R., Windschitl, P. D., & Suls, J. (2003). Egocentrism, event frequency, and comparative optimism: When what happens frequently is "more likely to happen to me" *Personality and Social Psychology Bulletin*, *29*, 1343–1356.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, *72*, 193–204.

Chowdhury, R., Sharot, T., Wolfe, T., Düzel, E., & Dolan, R. J. (2013). Optimistic update bias increases in older age. *Psychological Medicine*, 4, 1-10. [Epub ahead of print].

Christensen-Szalinski, J. J., & Beach, L. R. (1984). The citation bias: Fad and fashion in the judgment and decision literature. *American Psychologist*, *39*, 75–78.

Coady, C. A. J. (1992). *Testimony: A philosophical study*. Oxford: Oxford University Press.

Corner, A., & Hahn, U. (2013). Normative theories of argumentation: Are some norms better than others? *Synthese*, *190*, 3579–3610.

Corner, A., Harris, A., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 1625–1630). Austin, TX: Cognitive Science Society.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Covey, J. A., & Davies, A. D. M. (2004). Are people unrealistically optimistic? It depends how you ask them. *British Journal of Health Psychology*, *9*, 39–49.

Cox, R. (1946). Probability frequency and reasonable expectation. *American Journal of Physics*, *14*, 1–13.
Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin*, *52*, 177–193.
Crupi, V., Tentori, K., & Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, *116*, 971–985.
Dai, X., Wertenbroch, K., & Brendl, C. M. (2008). The value heuristic in judgments of relative frequency. *Psychological Science*, *19*, 18–20.
Dawes, R. M. (1976). Shallow psychology. In J. S. Carroll, & J. W. Payne (Eds.), *Cognition and social behavior*. Oxford, UK: Lawrence Erlbaum.
DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology*, *56*, 196–207.
De Finetti, B. (1974). *Theory of Probability*, (Vol. 1). NewYork: Wiley.
Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology*, *75*, 53–69.
Doherty, M., Mynatt, C., Tweney, R., & Schiavo, M. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*, 111–121.
Domingos, P. (2000). A unified bias-variance decomposition and its applications. In *Proceedings of the seventeenth international conference on machine learning* (pp. 231–238). Stanford, CA: Morgan Kaufmann.
DuCharme, W. M. (1970). Response bias explanation of conservative human inference. *Journal of Experimental Psychology*, *85*, 66–74.
DuCharme, W., & Peterson, C. (1968). Intuitive inference about normally distributed populations. *Journal of Experimental Psychology*, *78*, 269–275.
Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*, 69–106.
Dymond, R. F. (1949). A scale for the measurement of empathic ability. *Journal of Consulting Psychology*, *13*, 127–133.
Dymond, R. F. (1950). Personality and empathy. *Journal of Consulting Psychology*, *14*, 343–350.
Eagly, A. H., Chen, S., Chaiken, S., & Shaw-Barnes, K. (1999). The impact of attitudes on memory: An affair to remember. *Psychological Bulletin*, *125*, 64–89.
Eagly, A. H., Kulesa, P., Brannon, L. A., Shaw, K., & Hutson-Comeaux, S. (2000). Why counterattitudinal messages are as memorable as proattitudinal messages: The importance of active defense against attack. *Personality and Social Psychology Bulletin*, *26*, 1392–1408.
Eagly, A. H., Kulesa, P., Chen, S., & Chaiken, S. (2001). Do attitudes affect memory? Tests of the congeniality hypothesis. *Current Directions in Psychological Science*, *10*, 5–9.
Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processes. *Journal of Mathematical Psychology*, *2*, 312–329.
Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.
Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.
Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, *71*, 5–24.
Eil, D., & Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*, 114–138.
Elqayam, S., & Evans, J. St. B. T. (2011). Subtracting 'ought' from 'is': Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*, 233–248.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and under-confidence: The role of error in judgement processes. *Psychological Review*, *101*, 519–527.

Fiedler, K., & Kareev, Y. (2006). Does decision quality (always) increase with the size of information samples? Some vicissitudes in applying the law of large numbers. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, *32*, 883–903.

Fiedler, K., & Krueger, J. I. (2011). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social judgment and decision making*. New York, Taylor and Francis: Psychology Press.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239–260.

Frable, D. E. S. (1993). Being and feeling unique: Statistical deviance and psychological marginality. *Journal of Personality*, *61*, 85–110.

Funder (2000). Gone with the wind: Individual differences in heuristics and biases undermine the implication of systematic irrationality. *Behavioral and Brain Sciences*, *23*, 673–674.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75–90.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670.

Gage, N. L., & Cronbach, L. J. (1955). Conceptual and methodological problems in interpersonal perception. *Psychological Review*, *62*, 411–422.

Gage, N. L., Leavitt, G. S., & Stone, G. C. (1956). The intermediary key in the analysis of interpersonal perception. *Psychological Bulletin*, *53*, 258–266.

Galesic, M., Olsson, H., & Rieskamp, J. (2012). Social sampling explains apparent biases in judgments of social environments. *Psychological Science*, *23*, 1515–1523.

Galesic, M., Olsson, H., & Rieskamp, J. (2013). False consensus about false consensus. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 472–476). Austin, TX: Cognitive Science Society.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases" In W. Stroebe, & M. Hewstone (Eds.), *European review of social psychology: Vol. 2*, (pp. 83–115). Chichester, England: Wiley.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

Gigerenzer, G. (2006). Surrogates for theories. *Theory & Psychology*, *8*, 195–204.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge, UK: Cambridge University Press.

Gigerenzer, G., Todd, P. M., & The ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology*, *44*, 1110.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Gordon, R., Franklin, N., & Beck, J. (2005). Wishful thinking and source monitoring. *Memory and Cognition*, *33*, 418–429.

Green, P. W., Halbert, M. H., & Minas, J. S. (1964). An experiment in information buying. *Advertising Research*, *4*, 17–23.

Green, P. E., Halbert, M. H., & Robinson, P. J. (1965). An experiment in probability estimation. *Journal of Marketing Research*, *2*, 266–273.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
Gregg, A. P., & Sedikides, C. (2004). Is social psychological research really so negatively biased? *Behavioral and Brain Sciences*, *27*, 340.
Griffin, D., Gonzalez, R., & Varey, C. (2001). The heuristics and biases approach to judgment under uncertainty. *Blackwell Handbook of Social Psychology: Intra-Individual Processes*: *Vol. 1*, (pp. 207–235).
Hahn, U. (2014). Experiential limitation in judgment and decision. *Topics in Cognitive Science*. http://dx.doi.org/10.1111/tops.12083. [Epub ahead of print].
Hahn, U., Harris, A. J. L., & Corner, A. J. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*, 337–367.
Hahn, U., Oaksford, M., & Harris, A. J. (2012). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 15–38). Netherlands: Springer.
Harris, A. J. L., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition*, *110*, 51–64.
Harris, A. J. L., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*, 1366–1373.
Harris, A. J. L., & Hahn, U. (2011). Unrealistic optimism about future life events: A cautionary note. *Psychological Review*, *118*, 135–154.
Harris, A. J. L., Hsu, A. S., & Madsen, J. K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using ad hominem. *Thinking and Reasoning*, *18*, 311–343.
Harris, A. J. L., & Osman, M. (2012). The illusion of control: A Bayesian perspective. *Synthese*, *189*(1, Suppl. 1), 29–38.
Harris, A. J. L., Shah, P., Catmur, C., Bird, G., & Hahn, U. (2013). Autism, optimism and positive events: Evidence against a general optimistic bias. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 555–560). Austin, TX: Cognitive Science Society.
Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*, 555–588.
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.
Haselton, M. G., & Funder, D. C. (2006). The evolution of accuracy and bias in social judgment. In M. Schaller, et al. (Eds.), *Evolution and social psychology* (pp. 15–37). New York: Psychology Press.
Hastie, R., & Rasinski, K. A. (1988). The concept of accuracy in social judgment. In D. Bar-Tal & A. W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 193–208). Cambridge, England: Cambridge University Press.
Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.
Hertwig, R., & Pleskac, T. J. (2008). The game of life: How small samples render choice simpler. In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 209–235). Oxford, England: Oxford University Press.
Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*, 225–237.
Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, *138*, 211–237.

Hilton, D. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, *118*, 248–271.
Horvitz, E. J., Heckerman, D., & Langlotz, C. P. (1986). A framework for comparing alternative formalisms for plausible reasoning. In *Proceedings of the 5th national conference on AI (AAAI-1986)* (pp. 210–214).
Howson, C., & Urbach, P. (1996). *Scientific reasoning: The Bayesian approach* (2nd edition). Chicago, IL: Open Court.
Jarvstad, A., & Hahn, U. (2011). Source reliability and the conjunction fallacy. *Cognitive Science*, *35*(4), 682–711. http://dx.doi.org/10.1111/j.1551-6709.2011.01170.x.
Jarvstad, A., Hahn, U., Rushton, S., & Warren, P. (2013). Perceptuo-motor, cognitive and description-based decisions seem equally good. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 16271–16276.
Jarvstad, A., Hahn, U., Warren, P., & Rushton, S. (2014). Are perceptuo-motor decisions really more optimal than cognitive decisions? *Cognition*, *130*, 397–416.
Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*, 1–24.
Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*, 856–874.
Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences*, *23*, 681–683.
Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*: Cambridge: Cambridge University Press.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.
Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*, 123–141.
Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*, 582–591.
Kelley, H. H. (1950). The warm-cold variable in first impressions of persons. *Journal of Personality*, *18*, 431–439.
Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology*, *31*, 457–501.
Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, *102*, 390–402.
Klar, Y., Medding, A., & Sarel, D. (1996). Nonunique invulnerability: Singular versus distributional probabilities and unrealistic optimism in comparative risk judgments. *Organizational Behavior and Human Decision Processes*, *67*, 229–245.
Klayman, J., & Ha, Y. (1989). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.
Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1–53.
Korn, C., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*, 579–592.
Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological Bulletin*, *133*, 95–121.
Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, *27*, 313–327.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, *77*, 221–232.
Kruger, J., & Burrus, J. (2004). Egocentrism and focalism in unrealistic optimism (and pessimism). *Journal of Experimental Social Psychology*, *40*, 332–340.
Kruger, J., & Savitsky, K. (2004). The "reign of error" in social psychology: On the real versus imagined consequences of problem-focused research. *Behavioral and Brain Sciences*, *27*, 349–350.
Kruger, J., Windschitl, P. D., Burrus, J., Fessel, F., & Chambers, J. R. (2008). The rational side of egocentrism in social comparisons. *Journal of Experimental Social Psychology*, *44*, 220–232.
Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*, 395–409.
Kruglanski, A. W., & Ajzen, I. (1983). Bias and error in human judgment. *European Journal of Social Psychology*, *13*, 1–44.
Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.
Landahl, H. D. (1939). A contribution to the mathematical biophysics of psychophysical discrimination II. *The Bulletin of Mathematical Biophysics*, *1*, 159–176.
Leitgeb, H., & Pettigrew, R. (2010a). An objective justification of Bayesianism I: Measuring inaccuracy. *Philosophy of Science*, *77*, 201–235.
Leitgeb, H., & Pettigrew, R. (2010b). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*, 236–272.
Lench, H. C. (2009). Automatic optimism: The affective basis of judgments about the likelihood of future events. *Journal of Experimental Psychology. General*, *138*, 187–200.
Lench, H. C., & Bench, S. W. (2012). Automatic optimism: Why people assume their futures will be bright. *Social and Personality Psychology Compass*, *6*, 347–360.
Lench, H. C., & Ditto, P. H. (2008). Automatic optimism: Biased use of base rate information for positive and negative events. *Journal of Experimental Social Psychology*, *44*, 631–639.
Lenski, G. E., & Leggett, J. C. (1960). Caste, class, and deference in the research interview. *American Journal of Sociology*, *65*, 463–467.
Levine, J. M., & Murphy, G. (1943). The learning and forgetting of controversial material. *The Journal of Abnormal and Social Psychology*, *38*(4), 507–517.
Levinstein, B. A. (2012). Leitgeb and Pettigrew on accuracy and updating. *Philosophy of Science*, *79*, 413–424.
Lewis, D. (1986). A subjectivist's guide to objective chance. In *Philosophical papers*: *Vol. 2*. (pp. 83–132). London: Oxford University Press.
Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*, 159–183.
Lindley, D. V. (1982). Scoring rules and the inevitability of probability. *International Statistical Review*, *50*, 1–26.
Lindley, D. (1994). Foundations. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 3–15). Chichester, UK: John Wiley & Sons.
Lord, C., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.
Macdougall, R. (1906). On secondary bias in objective judgments. *Psychological Review*, *13*, 97–120.
Mandel, D. R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, *106*, 130–156.

Manz, W. (1970). Experiments on probabilistic information processing. *Acta Psychologica*, *34*, 184–200.
McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review*, *90*, 215–238.
Meacham, C. J. G. (2010). Two mistakes regarding the principal principle. *The British Journal for the Philosophy of Science*, *61*, 407–431.
Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*, 213–225.
Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). *Managing self-confidence*: *Theory and experimental evidence*. Working paper.
Moore, D. A. (2007). Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organizational Behavior and Human Decision Processes*, *102*, 42–58.
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*, 502–517.
Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgment: on being both better and worse than we think we are. *Journal of Personality and Social Psychology*, *92*(6), 972–989. http://dx.doi.org/10.1037/0022-3514.92.6.972.
Moore, D., & Small, D. (2008). When it is rational for the majority to believe that they are better than average. In J. I. Krueger (Ed.), *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes* (pp. 141–174). New York, NY: Psychology Press.
Morlock, H. (1967). The effect of outcome desirability on information required for decisions. *Behavioral Science*, *12*, 296–300.
Moutsiana, C., Garrett, N., Clarke, R. C., Lotto, R. B., Blakemore, S. J., & Sharot, T. (2013). Human development of the ability to learn from bad news. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(41), 16396–16401.
Mullen, B., Dovidio, J. F., Johnson, C., & Copper, C. (1992). In-group-out-group differences in social projection. *Journal of Experimental Social Psychology*, *28*, 422–440.
Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*, 151–166.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.
Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology*. *General*, *138*, 517–534.
Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and short-comings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.
Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
Olsson, E. J. (2013). A Bayesian simulation model of group deliberation and polarization. In *Bayesian argumentation* (pp. 113–133). Netherlands: Springer.
Olsson, E. J., & Vallinder, A. (2013). Norms of assertion and communication in social networks. *Synthese*, *190*, 2557–2571.
Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A′ and other modern misconceptions about signal detection theory. *Psychonomic Bulletin and Review*, *10*, 556–569.
Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*, 29–46.
Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *IRE Professional Group on Information Theory*, *PGIT-4*, 171–212.

Peterson, C. R., & DuCharme, W. M. (1967). A primacy effect in subjective probability revision. *Journal of Experimental Psychology*, *73*, 61–65.
Peterson, C. R., DuCharme, W. M., & Edwards, W. (1968). Sampling distributions and probability revisions. *Journal of Experimental Psychology*, *76*(2 pt. 1), 236–243.
Peterson, C., & Miller, A. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, *70*, 117–121.
Peterson, C., Schnieder, R., & Miller, A. (1965). Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology*, *69*, 522–527.
Peterson, C., & Uleha, Z. (1964). Uncertainty, inference difficulty and probability learning. *Journal of Experimental Psychology*, *67*, 523–530.
Peterson, C., Uleha, Z., Miller, A., & Bourne, L. (1965). Internal consistency of subjective probabilities. *Journal of Experimental Psychology*, *70*, 526–533.
Phillips, L., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.
Pierce, A. H. (1901). *Studies in auditory and visual space perception*. New York: Longmans.
Pitz, G. F. (1969a). The influence of prior probabilities on information seeking and decision-making. *Organizational Behavior and Human Performance*, *4*, 213–226.
Pitz, G. F. (1969b). An inertia effect (resistance to change) in the revision of opinion. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *23*, 24–33.
Pitz, G. F., Downing, L., & Reinhold, H. (1967). Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *21*, 381–393.
Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
Price, P. C., Pentecost, H. C., & Voth, R. D. (2002). Perceived event frequency and the optimistic bias: Evidence for a two-process model of personal risk judgments. *Journal of Experimental Social Psychology*, *38*, 242–252.
Radzevick, J. R., & Moore, D. A. (2008). Mypoic biases in competitions. *Organizational Behavior and Human Decision Processes*, *107*, 206–218.
Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.
Rosenkrantz, R. D. (1992). The justification of induction. *Philosophy of Science*, *59*, 527–539.
Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*, 279–301.
Ross, L., & Lepper, M. R. (1980). The perseverance of beliefs: Empirical and normative considerations. In R. A. Shweder (Ed.), *New directions for methodology of behavioral science fallible judgment in behavioral research* (pp. 17–36). San Francisco: Jossey-Bass.
Roy, M. M., Liersch, M. J., & Broomell, S. (2013). People believe that they are prototypically good or bad. *Organizational Behavior and Human Decision Processes*, *122*, 200–213.
Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance*, *27*, 153–196.
Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Evanston, IL: Northwestern University Press.
Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. New York, London: Psychology Press.
Shah, P., Harris, A. J. L., Bird, G., Catmur, C., & Hahn, U. (2013). *A pessimistic view of optimistic belief updating*. Manuscript under revision.
Sharot, T. (2012). *The optimism bias: Why we're wired to look on the bright side*. London, UK: Constable & Robinson Limited.

Sharot, T., Guitart-Masip, M., Korn, C. W., Chowdhury, R., & Dolan, R. J. (2012). How dopamine enhances an optimism bias in humans. *Current Biology*, *22*, 1477–1481.
Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., & Dolan, R. J. (2012). Selectively altering belief formation in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 17058–17062.
Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*, 1475–1479.
Shepperd, J. A., Klein, W. M. P., Waters, E. A., & Weinstein, N. D. (2013). Taking stock of unrealistic optimism. *Perspectives on Psychological Science*, *8*, 395–411.
Simmons, J. P., & Massey, C. (2012). Is optimism real? *Journal of Experimental Psychology. General*, *141*, 630–634.
Simon, H. A. (1978). Rationality as process and as product of thought. *The American Economic Review*, *68*, 1–16.
Slovic, P. (1966). Value as a determiner of subjective probability. *IEEE Transactions on Human Factors in Electronics*, *HFE-7*, 22–28.
Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, *6*, 649–744.
Snow, P. (1998). On the correctness and reasonableness of Cox's Theorem for finite domains. *Computational Intelligence*, *14*, 452–459.
Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137–149.
Stanovich, K., & West, R. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–726.
Suls, J., Wan, C. K., & Sanders, G. S. (1988). False consensus and false uniqueness in estimating the prevalence of health-protective behaviors. *Journal of Applied Social Psychology*, *18*, 66–79.
Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, *47*, 143–148.
Swann, W. B., Jr. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review*, *91*, 457–477.
Swets, J. A. (Ed.). (1964). *Signal detection and recognition by human observers: Contemporary readings*. New York: Wiley.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.
Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193–201.
Tetlock, P. E., & Levi, A. (1982). Attribution bias: On the inconclusiveness of the cognition-motivation debate. *Journal of Experimental Social Psychology*, *18*, 68–88.
Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*, 727–741.
Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*, 680–683.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.
Ulehla, Z. J. (1966). Optimality of perceptual decision criteria. *Journal of Experimental Psychology*, *71*, 564–569.
Vaughan, Wayland F. (1936). In: *General psychology* (pp. 211–237). Garden City, NY, USA: Doubleday, Doran & Company. http://dx.doi.org/10.1037/11466-007.? xxi, 634 pp.

von Mises, R. (1957/1981). *Probability, statistics and truth* (2nd rev. English ed.). New York: Dover.
Wald, A. (1950). *Statistical decision functions*. New York: Wiley.
Waldum, E. R., & Sahakyan, L. (2012). Putting congeniality effects into context: Investigating the role of context in attitude memory using multiple paradigms. *Journal of memory and language*, *66*(4), 717–730.
Walter, F. M., & Emery, J. (2006). Perceptions of family history across common diseases: A qualitative study in primary care. *Family Practice*, *23*, 472–480.
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140.
Wason, P. C. (1962). Reply to Wetherick. *Quarterly Journal of Experimental Psychology*, *14*, 250.
Wason, P. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.
Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*, 806–820.
Weinstein, N. D. (1982). Unrealistic optimism about susceptibility to health problems. *Journal of Behavioral Medicine*, *5*, 441–460.
Weinstein, N. D. (1984). Why it won't happen to me: Perceptions of risk factors and susceptibility. *Health Psychology*, *3*, 431–457.
Weinstein, N. D. (1987). Unrealistic optimism about susceptibility to health problems: Conclusions from a community-wide sample. *Journal of Behavioral Medicine*, *10*, 481–500.
Weinstein, N. D., & Klein, W. M. (1995). Resistance of personal risk perceptions to debiasing interventions. *Health Psychology*, *14*, 132–140.
Weinstein, N. D., & Klein, W. M. (1996). Unrealistic optimism: Present and future. *Journal of Social and Clinical Psychology*, *15*, 1–8.
Wendt, D. (1969). Value of information for decisions. *Journal of Mathematical Psychology*, *6*, 430–443.
West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, *118*, 357–378.
Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology*, *14*(4), 246–249.
Wheeler, G., & Beach, L. R. (1968). Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance*, *3*, 36–46.
Winterfeldt, von D., & Edwards, W. (1982). Costs and payoffs in perceptual research. *Psychological Bulletin*, *91*, 609–622.
Wolff, W. (1933). The experimental study of forms of expression. *Journal of Personality*, *2*, 168–176.
Wolpert, D. H. (1997). On bias plus variance. *Neural Computation*, *9*, 1211–1243.