CHAPTER 7

Cognitive Heuristics and Biases

(cowritten with Joshua Mugg)

> Man is not a rational animal, he is a rationalizing animal.
>
> – Robert Heinlein, *Tunnel in the Sky*

> I think unconscious bias is one of the hardest things to get at.
>
> – Ruth Bader Ginsburg

7.1 Introduction

This chapter focuses on heuristics and biases. A couple of preliminaries are in order. First, we are concerned with *cognitive* heuristics and biases rather than *social* or *implicit* bias. The phenomena we are interested in considering as examples of putative cognitive kinds involve competence in reasoning, inference, and decision-making. They concern cognitive tasks that deploy capacities for logical reasoning, inductive inference, probabilistic and statistical thinking, decision theory, and related norms of rational thought. Second, we take a *bias* to involve a deviation from the norms of rationality whereas a *heuristic* need not entail such a departure. We take the heuristic to be the underlying rule, process, or computation that may (or may not) result in a bias. This terminology agrees broadly with standard usage in cognitive science; for example, in an introduction to a seminal collection of papers, Gilovich and Griffin (2002, 3) define biases as "departures from the normative rational theory that served as markers or signatures of the underlying heuristics." Similarly, Stanovich, West, and Toplak (2016, 1110; original emphasis) write: "The term 'biases' refers to the systematic errors that people make in choosing actions and in estimating probabilities, and the term 'heuristic' refers to *why* people often make these errors – because they use mental shortcuts (heuristics) to solve many problems." Kahneman (2011, 7) says of participants in some of his experiments that "[t]he reliance on the heuristic caused predictable biases (systematic errors) in their predictions." Finally, even though he has a rather different position on the nature

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

and prevalence of bias, Gigerenzer (2018, 306) states: "A bias is a systematic discrepancy between the (average) judgment of a person or a group and a true value or norm." Moreover, in his "fast and frugal heuristics" research program, heuristics are held to be "efficient cognitive processes that ignore information" (Gigerenzer & Brighton 2009, 107). Thus, there seems to be broad consensus in cognitive science that biases are systematic departures from norms of rationality, whereas heuristics are neutral rules that may or may not lead to bias, depending on the context of their deployment. In addition, for many of these researchers, heuristics are the underlying rules or principles from which cognitive biases stem. To be sure, some philosophers have used these terms somewhat differently. Antony (2016, 161; original emphasis) considers that "bias is an essential element in human epistemic success," and holds that "bias plays a *constructive* role in the development of human knowledge; it is an enabling condition of human cognitive achievement." In what follows, we will use the terminology that is more prevalent among cognitive scientists, not because we think that philosophers should always defer to scientific practice, but because it preserves a useful distinction between a pattern of reasoning that results in systematic error (bias) and one that does not necessarily do so (heuristic). In the rest of this chapter, our focus will be on heuristics rather than biases because we take heuristics to be more plausible candidates for kindhood than biases, as we shall explain, though we shall return to biases toward the end of the chapter. But before going on to focus on heuristics, we will outline two reasons for thinking that *cognitive bias* as an overarching category is not a promising candidate for being a real kind in cognitive science.

Cognitive biases seem to share a few common features. In addition to constituting departures or deviations from norms of rationality, they are also generally thought to be systematic errors in reasoning rather than occasional mistakes (as some of the above characterizations confirm). They are also considered to be widely shared among humans rather than idiosyncratic quirks found in a small number of individuals, though they are by no means always universal. Finally, it is often thought that they are hard to avoid, though many of them can be overcome with some instruction or by adopting debiasing strategies. But these additional commonalities (systematicity, prevalence, and relative unavoidability) are features found in many aspects of our psychological makeup and are not distinctive of cognitive bias (as opposed to, say, perceptual illusions), even when conjoined.[1]

[1] Additionally, although some cognitive biases are thought to be innate features of human cognition, many may not be innate but learned. Moreover, while some cognitive biases emerge more readily

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

Hence, it would seem as though the only feature that sets cognitive biases apart is indeed that they are departures from rationality. But if this feature is to serve as a property common to all cognitive biases, we need to get clearer on the characterization of such departures. That obviously requires saying something about rationality. In the context of the cognitive science of reasoning and decision-making, rationality is thought to consist in a broad set of diverse norms, ranging from rules of logic, to rules of probabilistic and statistical inference, to those of decision theory. The deviations from the norms in these cases have very different natures: from those that constitute errors in logic to those that are considered to be errors in maximizing utility. Given the diversity that these departures represent and the different effects that they have on the actions and utterances of human agents, there does not seem to be much prospect for unifying them on a causal basis. Cognitive bias is unlikely to be a single kind of causal disposition or process with unified effects.

Moreover, the category of cognitive bias is almost certainly not a single etiological kind that has a common causal origin or history. Various researchers have speculated that biases generally have a variety of origins in the human mind, though there is considerable debate about the cognitive underpinnings of those origins. To mention just a few such proposed causal origins: lack of cognitive resources or "mindware gaps" (Stanovich, Toplak, & West 2008), "cognitive miserliness" (Toplak, West, & Stanovich 2011), motivational factors (Oreg & Bayazit 2009), or failure to inhibit intuitive responses (De Neys 2010). Though there is no consensus on which set of causal factors lead to cognitive biases, there is near unanimity among researchers that biases have multiple causes and can issue from various different aspects of the human psychological and cognitive makeup. Hence, there is no prospect of etiological unification when it comes to the category of cognitive bias.[2]

There is another problem with considering cognitive bias to be a cognitive kind. It follows from the characterization of biases and heuristics that we have adopted, which is prevalent in contemporary cognitive science, that heuristics are a more fundamental feature of human cognition or

when subjects are under cognitive load or under time constraints, others manifest even under optimal conditions.

[2] It may be useful here to invoke the analogy between cognitive biases and perceptual illusions, which has been deployed by numerous authors (cf. Kornblith 1994; Stein 1997; Kruglanski & Gigerenzer 2011; Pohl 2017). This analogy is apt in various ways, including the lack of a unifying causal basis in both cases. Even if we restrict ourselves to visual illusions, it is clear that they have multiple causal origins, ranging from basic physiological mechanisms to top-down interference from higher cognitive processes.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

cognitive architecture than biases. While a biased response will be one possible effect of the heuristic, as we have seen, there may well be instances in which a heuristic manifests in veridical or rational responses as well. Thus, the heuristic will necessarily be more causally connected than the bias, since the bias is an effect of the heuristic, and the bias is related to underlying cognitive architecture via the heuristic. Therefore, we will proceed to examine the prospects for considering *cognitive heuristic* (or *heuristic* for short) to be a cognitive kind. After having examined the prospects for heuristics in Section 7.2, we will go on to look at the case of a more specific class of heuristics (*cognitive miserliness*) in Section 7.3. After concluding that neither the class of heuristics as a whole nor that particular subtype are good candidates for cognitive kinds, in Section 7.4, we will examine a yet more specific heuristic and its resultant bias (*myside* or *confirmation bias*), finding it to be a better candidate for a cognitive kind.

7.2 Heuristic as a Cognitive Kind

Although empirical work on heuristics and biases has proliferated for at least half a century and it now represents a significant and growing research program in cognitive science, there has not been much explicit attention devoted to the question whether heuristics constitute a cognitive kind. This question can be broken down into three separate questions:

a) Do all (or nearly all) heuristics collectively constitute a kind?
b) Do heuristics cluster in subtypes, such that one or more of these subtypes separately constitutes a kind?
c) Are there individual heuristics that are kinds?

We will consider these questions in this order, focusing on the first question in this section and the second and third in subsequent sections.

An early use of the term "heuristic" in cognitive science occurs in the classic paper by Newell and Simon (1976), "Computer Science as Empirical Inquiry: Symbols and Search," which is considered one of the founding documents of the field. There, Newell and Simon introduced the notion of a "Physical Symbol System" that "exercises its intelligence in problem solving by search" (1976, 120). The process that they called "heuristic search" is one whereby the system generates and progressively modifies symbol structures until it produces a solution to the cognitive problem. Crucially, the search is not an exhaustive one because such a system has limited processing resources; indeed, they emphasize that the resources are "scarce relative to the complexity of the situations with which they are

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

confronted" (Newell & Simon 1976, 120). Although this formulation is obviously vague, the key is that intelligent problem-solving always involves some selectivity rather than exhaustivity in the search for solutions. This proposed feature of heuristics, that they are *selective* cognitive processes, appears to be prevalent in many subsequent accounts in cognitive science. Moreover, a very common way to understand selectivity in this context is that it involves ignoring or omitting some information, as mentioned in Section 7.1. A closely related characterization is that heuristics solve problems by substituting a difficult problem with a simpler one that admits of an easier solution (Kahneman & Frederick 2002; Stanovich, Toplak, & West 2008, 263). The relation between the two formulations is not hard to find: One way of ignoring information is to substitute a complex problem with a simpler one, perhaps one that involves fewer variables or requires less processing. On this common understanding, heuristics are thought to be cognitive processes that perform well in certain contexts, despite the fact that they do not take into account all relevant information. This characterization would also seem to be consistent with some of the figurative characterizations of heuristics that are widely deployed, such as mental "shortcuts" or "rules-of-thumb," or colorful epithets that are applied to heuristics, such as "quick and dirty" (Gilovich & Griffin 2002) or "fast and frugal" (Gigerenzer 2004).

On its own, this characterization of a heuristic as *a cognitive process that ignores information* does not appear sufficient to establish it as a cognitive kind. We would need to know something further about the causal profile of such processes in order to determine whether this central feature of heuristics is either due to a common causal mechanism or issues in certain stable effects. But the prospects on either count are not promising. To see this, we will look at a couple of theoretical accounts of heuristics in cognitive science. The two most prominent rival accounts of the nature of heuristics are the one associated with dual-system theory and that posited by the research program of ecological rationality. In a "dual-system" or "dual-process" account of cognitive architecture, heuristics are generally considered to pertain to System 1, a collection of cognitive module-like systems that are supposed to have some common characteristics.[3] On this view, or family of views, there are two types of systems or processes that underlie human reasoning.

[3] Though there are substantive differences between dual-system and dual-process models, for our purposes here the differences do not matter greatly. On the former view, heuristics can be thought of as processes that are implemented by the systems, while on the latter view they are identical with such processes or at least some of them. In this section, we put things mainly in terms of dual systems, but what we say can be rephrased in terms of dual processes.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

Broadly speaking, System 1 (*S1*) is fast, automatic, and associative, while System 2 (*S2*) is slow, controlled, and rule based. On early versions of these theories, heuristics were thought to pertain exclusively to *S1* and were held to be the default processes of human reasoning, which can be overridden or corrected by processes issuing from *S2* (e.g. Evans 1989). On many such views, heuristics from *S1* often yield valid responses when it comes to problems that the human mind has been adapted to solve. But in contexts that are far removed from the adaptive environment, they can give rise to inaccurate or mistaken solutions to problems. In such contexts, they need to be overridden by inferential or decision-making processes from *S2*, so as not to lead us into error or generate biases. It would seem that on this view there is a cognitive commonality to all heuristics, namely that they all issue from *S1*. But that is not generally agreed by dual system theorists themselves. For example, in an early articulation of the view, Evans (1989) labeled the two systems as "heuristic" and "analytic," respectively, but in later incarnations of the theory, he emphasized that heuristics can be associated with *S2* as well as *S1*. Later on, Evans (2011, 93) observed that "cognitive biases are as often attributed to Type 2 as Type 1 processing," pointing out that heuristic processing may occur in both systems and that cognitive biases may stem from both (cf. Evans 2012).[4] This would also seem to be the view of Tversky and Kahneman (1974) and most researchers employing the dual-system approach. But even if we were to narrow down the category of heuristics by applying it only to rule-based cognitive processing in *S1* (as some early versions of dual-system theory did), we would need to identify what all *S1* processes have in common, and that is in turn a vexed question without a clear or settled answer. In fact, a strong case has been made that the sub systems or processes of *S1* are not unified in terms of their causal properties. Despite the fact that many dual-system theorists once proposed a "Standard Menu" of features shared by all *S1* sub systems or processes, these early accounts have largely been abandoned in the face of significant difficulties. Proponents and critics alike have pointed out that the properties or features characteristic of each of *S1* and *S2* actually crosscut one another rather than cluster in certain ways (see e.g. Samuels 2009b; Evans & Stanovich 2013; Mugg 2016). Thus, even if heuristics are understood as *S1* processes they do not seem to share certain distinct causal properties that would set them apart from other cognitive processes and issue in certain stable effects.

[4] Evans (2012, 16) writes: "But it is an error to think that Type 1 processing is necessarily biased or that Type 2 processing is necessarily logical and abstract … Both types of processing can lead to correct answers and both can lead to biases." Also: "with experience we may adopt quick and dirty heuristics which are still explicitly applied by Type 2 processing …" (Evans 2012, 23)

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

Meanwhile, on ecological rationality views, heuristics are "fast and frugal" processes that are designed to solve various adaptive problems (see e.g. Gigerenzer & Brighton 2009; Gigerenzer 2018). On such theories, heuristics are sometimes superior to algorithms that take into account all available information. It is not just that heuristics involve a trade-off between accuracy and efficiency, since in some environments "heuristics are more accurate than strategies that use more information and computation" (Gigerenzer & Brighton 2009, 116). Gigerenzer and colleagues provide a number of examples to support this seemingly paradoxical claim. Consider something like the recognition heuristic, according to which thinkers choose between two alternatives based on their recognition of one of the alternatives. To illustrate, if American students are given pairs of German cities and asked to choose the most populous one in each pair, they typically choose the one whose name they recognize, since they generally lack detailed knowledge about the cities. The recognition heuristic turns out to be highly successful in this task, for the simple reason that size and recognition are highly correlated, at least for American students and German cities. In general, the heuristic works if there is a correlation between recognition and criterion in the environment (Kruglanski & Gigerenzer 2011, 100). Therefore, proponents of ecological rationality tend to regard heuristics as being more efficient and less prone to error or bias than dual-system theorists. But they do not claim that heuristics are always or even predominantly efficient or error-free. Moreover, from this perspective, heuristics can be either intuitive or deliberate cognitive processes; indeed the very same heuristic can be deployed intuitively or deliberately (Kruglanski & Gigerenzer 2011). To be sure, Gigerenzer (2004, 63–64) characterizes heuristics as having three features in common: (1) They exploit evolved capacities, (2) they exploit structures of environments, and (3) they are distinct from optimization models. But the first two features are clearly not distinctive of heuristics since many, if not most, aspects of cognition exploit evolved capacities and the environment. As for the third feature, this follows directly from the fact that heuristics do not take into account the totality of information in solving problems.[5] Hence, it is safe to say that heuristics, on this ecological rationality view, have nothing unique in common apart from the fact that they are cognitive processes that ignore information. If we classify a cognitive process as a heuristic there is nothing more we can say

[5] This is not an objection to Gigerenzer, since he seems to put these forward as necessary conditions on heuristics rather than properties that are causally related to the central characteristic of heuristics.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

about it; there are no generalizations to be made beyond the one that we used to identify them in the first place.

This brief survey suggests that there is nothing common to cognitive processes that deploy or implement heuristics, beyond the fact that they ignore information, on either of the two major theoretical approaches to heuristics in cognitive science. In the following section, we will look at some recent attempts to further subdivide the category of heuristics into narrower categories, in order to determine whether there may be candidates for cognitive kinds among the subordinate categories of heuristics. In so doing, we will also further corroborate the claim that the category of heuristics itself does not seem to correspond to a cognitive kind.

7.3 Sub Categories of Heuristics as Cognitive Kinds

In the previous section, we considered whether the category *heuristic* corresponds to a cognitive kind. We concluded that, at least according to the dominant theoretical accounts of heuristics in the empirical literature, there was no feature that all heuristics shared beyond their being selective cognitive processes that do not take into account all relevant information. The question we need to consider in this section is whether any sub categories of heuristics might correspond to a cognitive kind. By focusing on what we take to be one of the most promising candidates, we will argue for an answer in the negative, at present, though we will conclude with a positive suggestion for researchers.

There is a certain family resemblance among some heuristics, such as vividness effects (e.g. representativeness bias), affect substitution, impulsively associative thinking, framing effects, anchoring effect, belief bias, denominator neglect, outcome bias, hindsight bias (also known as "curse of knowledge effects"), conjunction errors, and confirmation bias.[6] In fact, some look like instances, determinates, or subgroups of the others. For example, the anchoring effect looks like a framing effect pertaining to estimation of numbers when numbers are mentioned previously to the question. Likewise outcome bias, hindsight bias, belief bias, and confirmation bias look quite similar. There are, however, two problems with classifying distinct sub categories of heuristics by way of such family resemblance. First, as we shall see, there have been multiple attempts to categorize heuristics into subkinds, and they have not tended to match up very well. It seems that when we proceed by

[6] Although we label some of these as "biases" here, we do so only because that is what they are so called in the empirical literature. Again, we are interested in the heuristic underlying these biases.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

way of grouping heuristics together by mere similarity relations, there are too many ways to cut the cake, and several of the heuristics will fall into multiple categories. Second, a mere family resemblance is not enough for kindhood. In keeping with the account of cognitive kinds adopted in this book, we would need to understand *why* there is a family resemblance by establishing the *causal* connectedness of the shared properties. As of yet, such a principled causal basis for subdividing heuristics has yet to be provided.

This last claim can be further supported by taking a look at some proposed taxonomies of heuristics. We will argue that the nature of these categories also provides indirect support for the claim that there is no category, *heuristic*, characterized by a set of common features, that can be divided into subtypes that relate to it as species to genus. These taxonomies attempt to group heuristics and biases into a number of clusters based on their differentiating features. There have been a number of attempts to divide the domain of heuristics (and biases) into taxonomic categories, based on a variety of principles or theoretical considerations, but these sub categories are often disparate in terms of their causal and etiological profiles. Ceschi, Costantini, Sartori, et al. (2019) recently undertook an attempt to compare some of the prominent taxonomies in the cognitive science literature and tabulated the results for ease of comparison (see Figure 7.1).

There are several things to notice about this attempt to compare various taxonomies. First, most of these taxonomies do not distinguish clearly between heuristics and biases – though we might interpret this charitably to mean that they are interested in the heuristic underlying the bias in all cases. Second, different taxonomies deploy divergent categories to classify heuristics and biases and there is almost no overlap in the labels that they give to the categories that are deployed. Of course it may be that some categories as deployed by some theorists are just terminological variants of those adopted by other theorists (e.g. in Figure 7.1. Arnott's "Adjustment" category may correspond to Carter, Kaufmann, and Michel's "Reference Point" category, and Stanovich et al.'s "mindware gaps" may correspond to Oreg and Bayazit's "simplification biases"). But a little probing suggests that their categories only partially overlap and may even crosscut (e.g. Baron's "representativeness" and "availability" categories overlap Oreg and Bayazit's "simplification biases," which also includes phenomena not classified by Baron).[7] Third, and most importantly, even within each taxonomy, the categories are

[7] While there is nothing wrong with crosscutting categories in science (see Khalidi 2013), in this case, it does not appear that the different taxonomies are trying to capture different aspects of the heuristics (as when biologists classify organisms based on phylogenetic and ecological properties).

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

Heuristic / Bias \ *Source*	*Baron (2000)*	*Gilovich, Griffin, & Kahneman (2002)*	*Arnott (2006)*	*Carter, Kaufmann, & Michel (2007)*	*Stanovich, Toplak, & West (2008)*	*Oreg & Bayazit (2009)*
Gambler's fallacy	Representativeness	Representativeness and Availability	Statistical	Control illusion	Probability knowledge (Mindware gaps)	Simplification biases
Conjunction fallacy	Representativeness	Representativeness and Availability	Statistical	Control illusion	Probability knowledge (Mindware gaps)	Simplification biases
Representativeness heuristic	Representativeness	Representativeness and Availability	Statistical	Base rate	Probability knowledge (Mindware gaps)	Simplification biases
Base rate fallacy		Representativeness and Availability	Statistical	Base rate	Probability knowledge (Mindware gaps)	Simplification biases
Framing			Presentation	Presentation	Focal Bias	Regulation biases
Distinction bias						
Availability heuristic	Availability	Representativeness and Availability		Availability cognition	Mindware gaps	Simplification biases
Imaginability bias	Availability	Representativeness and Availability	Memory	Availability cognition	Mindware gaps	Simplification biases
Better than average effect				Output evaluation		Verification biases

Optimism bias	Effect of desire on belief	Optimism	Confidence	Output evaluation		Verification biases
Anchoring heuristic	Underadjustment	Anchoring contamination and Compatibility	Adjustment	Reference point	Focal bias	
Reference price			Adjustment	Reference point		
Regression toward the mean			Adjustment	Reference point		
Extra cost effect	Utility theory					
Sunk costs fallacy	Utility theory			Commitment		
Endowment effect	Diminishing sensitivity			Commitment		Regulation biases
Time discounting	Diminishing sensitivity					

Figure 7.1. Comparison of taxonomies of heuristics and biases from five different sources; parentheses indicates classification in more than one category (adapted from Ceschi, Costantini, Sartori, et al. 2019).

based on rather diverse theoretical considerations. This last point is perhaps the most significant for our purposes since it signals that there does not seem to be a common basis for classification even by a single group of theorists. For example, Oreg and Bayazit (2009) draw a tripartite distinction among biases based on the motivations that give rise to the biases (simplification, verification, regulation). Roughly speaking, simplification biases stem from a desire to achieve a comprehensible image of the world, while verification biases are motivated by the need to achieve consistency and coherence, and regulation biases arise from trying to approach pleasure and avoid pain. But they explicitly argue that there are complex direct and indirect relationships among the bias categories. For example, according to them, "verification biases contribute to the creation of regulation biases" (Oreg & Bayazit 2009, 189). Hence, the underlying dispositions are not independent of one another and do not divide the biases into disjoint categories. Similarly, Stanovich, Toplak, and West (2008) provide a taxonomy of heuristics and biases that is based on the nature of the breakdown in reasoning that results in the bias (e.g. cognitive miserliness, mindware gap, contaminated mindware), but they also acknowledge that some biases may belong to more than one category. Setting aside the fact that they lump heuristics and biases together, what this suggests is that some of the heuristics that they have identified are due to some general features of cognitive processing (e.g. cognitive miserliness) while others are just a result of a lack of cognitive skill or training (e.g. mindware gap). Finally, with regards to the kindhood of the superordinate category *heuristic*, none of these taxonomies identifies any common traits of heuristics – beyond ignoring information – that would unify them and enable us to distinguish their sub categories in the manner of genus and differentia. When one looks at the categories within each taxonomy, let alone across taxonomies, they do not seem to have any obvious shared features that would identify them all as subordinate categories of the superordinate category, *heuristic*. Thus, these taxonomies also provide indirect support for the claim that *heuristic* is not a homogeneous category, since its subordinate categories are not characterized by a common set of properties (beyond ignoring information).

This last conclusion is further corroborated by what seems to be the most comprehensive and "evidence-based" approach to establish a taxonomy of heuristics and biases. In arriving at their own taxonomy, Ceschi, Costantini, Sartori, et al. (2019) build partly on existing taxonomies (tabulated above) but they also pursue a strategy that relies on finding correlations between the performance of subjects on different cognitive tasks associated with heuristics and biases. Using a large number of participants

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

(*n* = 289) and a within-subjects design, they attempt to discern patterns of correlation in performance on a battery of seventeen cognitive tasks. Their method involves complex statistical techniques for discerning patterns among these tasks, but the method can be divided into two main steps. First, they performed a Multiple Correspondence Analysis (MCA) to determine the presence of common categories deployed in existing taxonomies. Then they performed a Principle Component Analysis (PCA) to assess relationships between biases belonging to the dimensions extracted from the MCA. This yielded three factors that were interpreted as the main categorical distinctions between types of heuristics or biases: (1) mindware gaps, (2) valuation biases, and (3) anchoring and adjusting. Here again, there is no suggestion that all heuristics have commonalities beyond ignoring information. The subordinate categories are not identified as species of a genus, each of which is characterized by a number of common properties in addition to certain distinguishing characteristics that differentiate them from other members of the genus. These considerations confirm the heterogeneous nature of the category of heuristics. But although this taxonomy, like the others already considered, does not give us any reasons for thinking that *heuristic* is a valid cognitive kind, it is possible that some of these subordinate categories correspond to cognitive kinds, not as species of a single genus but as stand-alone kinds. The specific categories that Ceschi, Costantini, Sartori, et al. (2019) identify do not seem to be promising candidates, as can be surmised by considering each very briefly. *Mindware gaps*, posited by Stanovich (2010), are instances in which thinkers simply lack the relevant reasoning principles to perform a cognitive task, as when they commit the conjunction fallacy, gambler's fallacy, or base rate fallacy. These supposed gaps in knowledge are clearly cognitively variegated and have nothing in common apart from being instances of ignorance of certain rules or principles. As for *valuation biases*, they include such biases as optimism bias, temporal discounting, and sunk cost fallacy, and appear to involve either over- or under-valuing certain outcomes. As Ceschi, Costantini, Sartori, et al. (2019, 197) admit, those who are susceptible to such biases can have opposing character traits (e.g. optimism and pessimism), which also suggests heterogeneity of the bias. Similarly, *anchoring and adjustment*, which includes framing effects and regression to the mean, seems to stem from a variety of traits or dispositions and is not correlated with other cognitive features. We will therefore take a closer look at another prominent sub category of heuristics, which is identified with its role in a specific cognitive architecture, to determine whether it might correspond to a cognitive kind in its own right.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

If cognitive kinds are individuated in terms of their causal role, a promising approach to distinguishing a sub category of heuristics is to do so with reference to a causal model of cognitive architecture. We have argued that heuristics all involve ignoring information or taking some sort of short-cut, but it is plausible that not all information ignoring or short-cutting are the same. It may be possible to identify various points in cognitive processing where information is ignored and identify corresponding heuristics. We might find patterns among the various individual heuristics based on the point at which information is ignored within cognitive processing. This would provide us with a taxonomy of heuristics, subdividing the category *heuristic* into genuine kinds, with each sub-kind having a distinct causal profile in human cognitive architecture. Such a taxonomy would provide us with an explanation for why the heuristics are so divided by relating each subdivision to the causal structure of cognitive processing. It would also aid in cases where it seems a heuristic fits into more than two categories: When specific heuristics seem to fit into more than one of these categories, we would have good reason to split the heuristic, as it would be implicated at two distinct points in reasoning processing. Some of the taxonomies of heuristics already mentioned seem to be attempting to divide heuristics along these lines. For example, Stanovich offers a framework for conceptualizing individual difference, which fits naturally with his taxonomy for classifying heuristics. This framework can provide the basis for identifying certain subtypes of heuristics that are likely to be candidates for cognitive kinds. The plausibility of sub-groupings of heuristics being kinds depends upon which category we consider and the cognitive architecture in which it is situated. It is not possible to run through every sub category that has been mentioned in the literature, but we can at least consider one of the more promising sub categories proposed in one of the taxonomies already mentioned. Stanovich claims that there is a group of heuristics that are unified because they all arise from *cognitive miserliness,* which itself arises, in part, because of time and computational restraints. In the remainder of this section, we will consider whether *cognitive miserliness*, as proposed by Stanovich, is a cognitive kind capable of unifying some of the items in the above list of individual heuristics.

In an attempt to describe cognitive miserliness, Stanovich, Toplak, and West (2008) break it down into two aspects or "rules." The first rule is: "Default to Type 1 processing whenever possible." This rule is conceived as a structural feature of our cognitive architecture and it presupposes a dual-process view of cognition. As such, its fortunes are tied to those of dual-process theory: If it turns out that dual-process theory is not an apt

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

theoretical account of human cognition, then this first aspect of cognitive miserliness simply cannot be sustained in its current form. Setting aside worries about dual process theory outlined in Section 7.2, let us grant for the sake of argument that something like that theory is correct. If we accept this cognitive architecture (or something like it), the construct of cognitive miserliness would seem to be a second-order rule or process that governs the Type 1 processes posited by dual-process theory. Cognitive miserliness is a process whereby cognition defaults to Type 1 processing. Presumably, a basic feature of cognitive architecture is that Type 1 is the default processing that conserves effort and requires fewer resources. But, according to Stanovich, Toplak, and West (2008), this is not the only way in which cognitive miserliness features in cognitive architecture. The second aspect or rule of cognitive miserliness is activated when Type 1 processing will not yield a solution; at that point, the thinker relies on serial associative cognition with a focal bias (which is a Type 2 heuristic process). Stanovich (2010, 67) expresses the "basic idea" behind focal bias as follows: "… the information processor is strongly disposed to deal only with the most easily constructed cognitive model." He also writes that "There are less expensive kinds of Type 2 processing that we tend to fall back on when Type 1 mechanisms are not available for solving the problem" (Stanovich 2010, 63). This takes place particularly in novel situations from an evolutionary point of view, where there are no stimuli that trigger Type 1 processes. Again, this is a second-order rule or process rather than a case of first-order cognitive processing.

This account of cognitive miserliness is couched in a cognitive architecture that posits three hierarchically organized systems in the mind: (1) the Autonomous Mind (associated with Type 1 processing); (2) the Algorithmic Mind (associated with Type 2 processing, responsible for cognitive ability measured by intelligence tests); and (3) the Reflective Mind (also associated with Type 2 processing, responsible for different cognitive styles) (Stanovich 2010, 35). This scheme is depicted in diagrammatic form in Figure 7.2. In this boxology, cognitive miserliness can be identified with two distinct processes. The first is represented by the horizontal arrow leading directly from the Autonomous Mind (responsible for Type 1 processing) to a response. The second is represented by the arrow labeled "E," leading from the Algorithmic Mind to response or attention. This suggests that cognitive miserliness is manifested in two entirely different cognitive processes or sub processes, pertaining to distinct psychological systems or capacities. In one case, it corresponds to a default to Type 1 processing, in another a resort to associative cognition (which is a Type 2

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

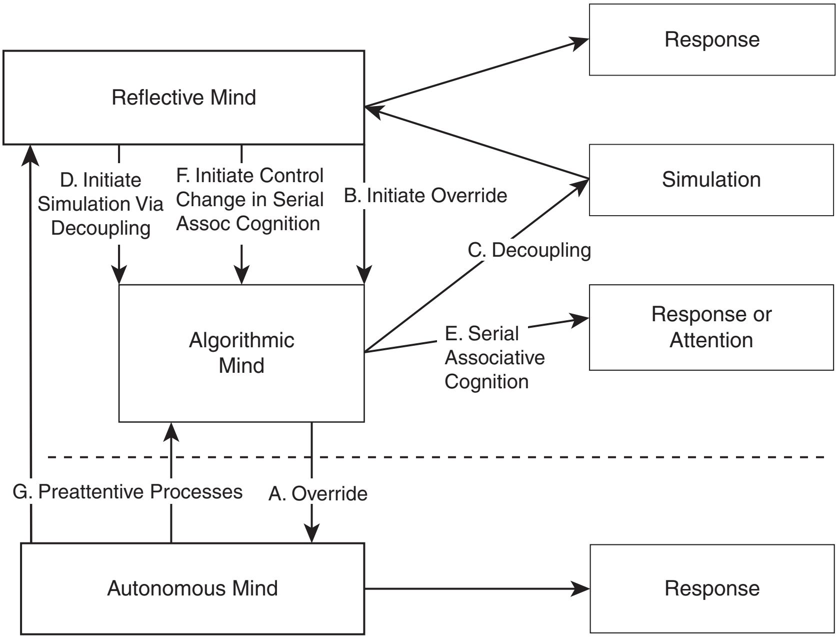
Response
Reflective Mind
Simulation
D. Initiate Simulation Via Decoupling
F. Initiate Control Change in Serial Assoc Cognition
B. Initiate Override
C. Decoupling
Response or Attention
Algorithmic Mind
E. Serial Associative Cognition
G. Preattentive Processes
A. Override
Autonomous Mind
Response

Figure 7.2. Tripartite model of the mind proposed by Stanovich (2010), showing the Autonomous Mind, Algorithmic Mind, and Reflective Mind, and some of their interactions.

process). As such, it does not seem to be a single type of cognitive process when viewed in the context of a causal model of cognitive architecture.[8] Now, it is possible that there is a unitary mechanism behind these different outcomes, perhaps some aspect of cognitive architecture that ensures that (in many contexts), cognition takes a path of least cognitive effort. Toplak, West, and Stanovich (2011, 1283) write: “Humans are cognitive misers because their basic tendency is to default to heuristic processing mechanisms of low computational expense.” But if so, such an underlying cognitive tendency has yet to be described in any detail. Moreover, if it is a second-order rule that specifies when to use Type 1 and Type 2 processing, then it would seem to pertain not only to heuristics but to reasoning or cognition more generally.

[8] In fact, Stanovich sometimes mentions “override failure” as a “third category” of cognitive miserliness. He writes that “in override failure, cognitive decoupling does occur, but it fails to suppress the Type 1 processing of the autonomous mind” (Stanovich 2010, 100). This failure would seem to correspond to the absence of the arrows labeled “A” and “B” in Figure 7.2 in the performance of some cognitive task.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

In the face of this objection, defenders of cognitive miserliness might attempt to validate the construct by linking it to an operational test. Stanovich and collaborators have made the case that cognitive miserliness is subject to individual difference and is associated with a certain cognitive style. Rather than regarding it as a universal human trait that is uniform across individuals, they argue that empirical results show considerable variation among individuals in terms of their tendency to be cognitive misers. According to them, avoiding cognitive miserliness requires first detecting the inadequacy of the Type 1 response, then using Type 2 processing both to suppress the Type 1 response and to come up with a better alternative (Stanovich, Toplak, & West 2020, 1122; see also Figure 7.2 above). They hold that these abilities, which vary among individuals, are not measured on standard intelligence (IQ) tests, but are rather correlated with performance on the Cognitive Reflection Test (CRT), which consists of just three math questions that are fairly simple to solve, but also tempt experimental participants to offer an intuitive but incorrect answer.[9] Poor performance on the CRT is evidence of miserliness and good performance is evidence of the opposite. Unlike performance on standard intelligence tests or IQ tests, which is not perfectly correlated with performance on the full array of tasks in the heuristics and biases literature, performance on CRT is so correlated. It is, therefore, a more direct measure of cognitive miserliness. Toplak, West, and Stanovich (2011, 1284) put it thus:

> In short, the CRT is a measure of the tendency toward the class of reasoning error that derives from miserly processing. … Intelligence tests do not assess the tendency toward miserly processing in the way that the CRT does. … The CRT measures miserliness in action, so to speak. It is a direct measure of miserly processing rather than an indirect self-report indicator.

In subsequent work, Stanovich (2016) developed another, more extensive version of the test, the Comprehensive Assessment of Rational Thinking (CART), which is held to be a more accurate measure of cognitive miserliness. It might seem as though this would provide some corroboration of miserliness as a cognitive kind, since we have an instrument that is designed to measure it and assess the extent to which individuals are cognitive misers. But the existence of such a test is not a sufficient vindication of the existence of the kind. If we want to know what cognitive miserliness

[9] One of these is the notorious "bat-and-ball problem": A bat and ball together cost 10.10. The bat costs a dollar more than the ball. How much does the ball cost? The correct answer is: five cents (since 11.05 + 12.05 = 13), but many participants give the tempting answer: ten cents (see Kahneman & Frederick 2002).

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

is, we can say that it is whatever is measured by a certain test, but then if we want to know what that test measures, it seems that the only answer we have available is that it measures cognitive miserliness. The name of the revised test suggests that there is an independent construct being measured, namely Rational Thinking, but as pointed out in Section 7.1, rationality is a heterogeneous category. More importantly, rationality is far broader than just a lack of cognitive miserliness,[10] as Stanovich and colleagues agree, since they also think that rationality involves such attributes as avoiding "mindware gaps." The existence of the test is not sufficient to validate cognitive miserliness as a construct, nor establish it as a cognitive kind.[11] Therefore, *cognitive miserliness* does not unify a sub group of heuristics into a kind.

If Stanovich's attempt to categorize heuristics into subordinate kinds using *cognitive miserliness* fails, that does not imply that the overall approach of using cognitive architecture to provide a taxonomy of heuristics is flawed. There are other cognitive architectures, and each may be able to provide its own taxonomy of heuristics. Indeed, it may even be that a dual-process taxonomy of heuristics could be developed apart from cognitive miserliness. This provides a possibly fruitful avenue for researchers interested in looking at the relation between individual heuristics: to develop taxonomies of heuristics based on the various cognitive architectures currently on offer, with distinct sub categories of heuristics corresponding to elements of cognitive processing. One might worry about waiting on a completed cognitive architecture to determine the kindhood of the sub categories of heuristics. After all, the existence of the various

[10] In some work, the aim is said to be to come up with a test of the "Rationality Quotient" (RQ), along the lines of the Intelligent Quotient (IQ) measured by intelligence tests (Stanovich 2010, 189–190). Other constructs have also been proposed, such as "active open-minded thinking," but in the absence of some independent account of what these constructs are, this does not get us out of the circle.

[11] Could Stanovich and collaborators claim that one *part* of the test measures cognitive miserliness? Stanovich (2016, 29) identifies four subtests within CART that test for avoiding miserly processing, but he says the following about most of these subtests: "All of these tasks and their associated effects, although involving miserly processing, are still quite complex tasks. More than miserly processing is going on when someone answers suboptimally in all of them." So it does not seem as if any one subtest is an operational test for miserly processing (or the avoidance thereof). The one test that he implies is most geared to miserly processing is the "Reflection versus Intuition subtest" and the task that he mentions is the famous "bat-and-ball" problem (see footnote 8). However, he does not say whether we should take this task as an operational test for miserly processing. Even if we were to consider this task (and perhaps others like it) as an operationalization of the category of miserly cognitive processing, then that would not be sufficient to show that the construct is a valid one. We would still need a characterization of cognitive miserliness that situates it within a causal network. As famously argued by Cronbach and Meehl (1955): One needs to have a valid construct rooted in a "nomological net" before one can proceed to operationalize it (see also Flake & Fried 2020).

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

heuristics is the explanandum for which cognitive architectures are built as explanations. This may result in a temporary impasse, but it may not. With taxonomies of heuristics from the various cognitive architectures in hand, we can compare similarities and differences. One possibility is that the various cognitive architectures, though differing in where ignoring information figures within the overall cognitive processing, will produce similar taxonomies, indicating that some heuristics cluster to constitute cognitive kinds. In that case, we need not first determine which cognitive architecture is correct in order to determine whether some of the heuristics cluster into kinds. Of course, it may turn out that the taxonomies do not match up very well. In the meantime, we should consider whether individual heuristics are good candidates for kindhood.

7.4 Confirmation Bias or Myside Heuristic

Confirmation bias is one of the earliest biases discussed in the literature on heuristics and biases. In this section, we will begin by examining the evidence for a confirmation bias, but we will also consider the case for the existence of a *heuristic* underlying the bias, for reasons provided in Section 7.1. (In what follows, we will talk mainly in terms of "confirmation bias," since that is the preferred term in the empirical literature, but our real focus is the putative heuristic causing the bias.) Is there a cognitive kind corresponding to the category of *confirmation bias*, and if so, what are its main features? Moreover, what is the relationship between confirmation bias and myside bias, and should the former be replaced by the latter, as suggested by some researchers (e.g. Mercier 2017)?

Some of the earliest experiments that purported to show a confirmation bias in human subjects were reported by Wason (1960) using the so-called 2-4-6 task. The results that he obtained were supposed to show that a significant number of experimental participants are "unable, or unwilling, to test their hypotheses" (1960, 129). Wason gave participants the sequence of numbers 2-4-6 and asked them to guess the simple rule to which the numbers conform. To this end, participants were asked to provide guesses of other number triples to the experimenter and the reason for each guess, at which point the experimenter would tell them either that their guess was an instance of the rule or not. Participants were told that once they felt "highly confident" that they had hit upon the right rule, they were supposed to make an "announcement." The sequence 2-4-6 might suggest that the rule is something like "even numbers" or "consecutive even numbers" or "ascending consecutive even numbers," but the correct rule that Wason

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

had in mind is just "ascending numbers." In the original experiment, six of twenty-nine participants (21 percent) guessed the correct rule on the first announcement, and ten (34 percent) guessed correctly on the second announcement. But the main finding was that those who did not get it right on the first or second announcement did not attempt to "test their hypotheses" in the sense of providing sequences that they did not think conformed to the rule, in order to rule out certain hypotheses. This has been taken to show that at least some people have a bias to confirm their hypotheses rather than disconfirm them (though Wason did not put it in these terms nor use the term "confirmation bias" in his original paper).

There are a number of things to notice about Wason's experiment and the conclusion that has often been drawn from it. First, a majority (55 percent) of participants performed quite well, hitting upon the correct rule on the first or second announcement and providing instances that were both compatible and incompatible with their hypotheses, thus effectively testing them (not just confirming them). Second, the task is a tricky one. As Wason (1960, 138) admits, one possible explanation of why some participants did not perform well (i.e. required more than two announcements) is that "the correct rule (increasing magnitude) was so trivial that students would have been reluctant to entertain it." Participants may not have provided enough negative instances because they did not consider any rules that are less specific than the obvious ones. Third, other psychologists have pointed out that this should not be seen as a decisive demonstration of a confirmation bias but rather a "positive test strategy," where a positive test strategy is one in which one tests cases that one thinks conform to one's hypothesis. In this case, if one's initial hypothesis is, "consecutive even numbers," then one would give instances conforming to it (e.g. 10-12-14, 98-100-102). In Wason's case, the true hypothesis was much broader than expected, so these guesses did not only conform to the participants' hypothesis, they were also in conformity with those of the experimenter. But if one's initial hypothesis is broader than the true hypothesis, or if it is overlapping, or disjoint, then proposing conforming instances can certainly lead to falsifying the hypothesis and suggesting alternatives (Klayman & Ha 1987; Klayman 1995). In general, adopting such a strategy need not produce systematic error, since one can discover that one's hypothesis is wrong by proceeding in this way, depending on the context. Hence, the experiment does not demonstrate a confirmation bias, though it may well show what could be labeled a "positive test heuristic."

In the decades since Wason's work, numerous researchers have pointed to cases that seem more like genuine instances of bias when it comes to

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

confirming and disconfirming hypotheses. Edwards and Smith (1996) posit a "disconfirmation bias" when it comes to beliefs that are contrary to one's own beliefs. The emphasis in this work is on cases in which people attempt to *undermine* evidence that is contrary to their beliefs, though on a plausible model of credence, decreasing credence in an incompatible belief increases credence in one's existing beliefs.[12] In one experiment, Edwards and Smith (1996) chose seven issues about which participants had strong prior beliefs (as determined in pretesting several weeks preceding the experiment), such as the death penalty and corporal punishment. They presented participants with two arguments on each issue, consisting of a single premise and conclusion, one defending a certain position and the other defending the opposite position. In the first stage, participants were asked to rate the strength of each argument, and in the second stage, they were asked to list all the thoughts that occurred to them when they considered the conclusions of each of the arguments. They found that individuals judged arguments supporting beliefs that are incompatible with their own beliefs to be weaker, they spent more time scrutinizing the arguments, they generated a greater number of relevant thoughts about them, and they produced a greater number of arguments refuting those arguments (Edwards & Smith 1996, 14). This is often regarded as a seminal study showing that people are generally biased against arguments that conflict with or undermine their own beliefs. The bias consists both in a judgment concerning the strength of the opposing argument and in the time and effort expended in refuting it. Therefore, this can be considered a *disconfirmation* bias as opposed to a confirmation bias – but one directed at incompatible or contrary beliefs. Based on this and similar work, some researchers think that it is misleading to talk about a confirmation bias, since it is more accurate to say that people have a "myside bias" (Mercier 2017), favoring evidence that supports their own beliefs and disfavoring evidence that weakens them. Moreover, we would argue that since there are circumstances in which such a cognitive tendency may not be irrational or violate norms of inference (as we shall see shortly), it should be considered instead a "myside heuristic," given the terminology that we have adopted in this chapter. Therefore, in the rest of this section, we will address the question whether the *myside heuristic* can be considered a cognitive kind.

[12] Compare Mercier and Sperber (2011, 64) on confirmation bias: "It is a bias in favor of confirming one's own claims, which should be naturally complemented by a bias in favor of disconfirming opposing claims and counterarguments."

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

The main obstacle to considering *myside heuristic* to be a cognitive kind is the apparent heterogeneity when it comes to the kinds of psychological phenomena that it comprises. Some researchers have pointed out that the psychological processes that have been identified as factors in this body of experimental work range from relatively low-level perceptual or attentional mechanisms to higher-level cognitive dispositions to interpret and evaluate evidence and generate hypotheses. Many of these tendencies can be considered heuristics in the sense of rules or procedures that ignore information, but it may seem unlikely, given what we know about cognitive architecture, that it would be the very same process that is operative in these apparently disparate domains. Moreover, if these phenomena are all confirmed, it would seem that they often push in opposite directions, as it were, sometimes tending to confirm hypotheses and at other times tending to disconfirm them, depending on whether they are one's own hypotheses or incompatible ones. Could there be a unifying underlying cause that is responsible for all or at least a significant portion of these phenomena?

Given this apparent diversity, it would seem more promising to focus on one of the various phenomena at issue, such as the attitude toward evidence or arguments supporting and opposing one's own belief or hypothesis. If there is such an attitude, it would conform to the characterization of a heuristic that we're operating with in this chapter, since it is likely to involve one or more cognitive processes that ignore relevant information, in this case, either the actual evidence against a belief, the strength of that evidence, or the relative strength of the evidence for and against that belief. Nickerson (1998, 178) characterizes it as follows: "the tendency to give greater weight to information that is supportive of existing beliefs or opinions than to information that runs counter to them." What grounds do we have for positing a version of myside heuristic according to which subjects ignore arguments or evidence against their own beliefs in favor of evidence for those beliefs, and could such a tendency constitute a cognitive kind? Mercier and Sperber (2011, 63) hold that, as they have characterized it, confirmation (or myside) bias is "one of the most studied biases in psychology …" and survey some of the research studies in its favor. One such study has already been summarized above (Edwards & Smith, 1996) and it clearly illustrates that people judge evidence or arguments supporting their own beliefs to be stronger than evidence or arguments that are incompatible with their beliefs. In another influential study, Lord, Ross, and Lepper (1979) selected forty-eight participants, evenly divided among "proponents" and "opponents" of capital punishment, as determined in a pre test questionnaire. Participants were shown (fictitious) research results

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

that either confirmed or disconfirmed their position, followed by detailed descriptions of the research procedure, along with critiques of the research and rebuttals by the supposed authors. All participants were exposed to both confirming and disconfirming information, counterbalanced to control for order effects. Asked for their final attitudes on the issue of capital punishment, relative to the experiment's start, proponents reported that they were more in favor of capital punishment, whereas opponents reported that they were less in favor (Lord, Ross, & Lepper 1979, 2103–2104). The researchers propose that when individuals encounter evidence that both supports and undermines one of their beliefs, they will assimilate the former while dismissing and discounting the latter. This "biased assimilation" of evidence in turn leads to belief polarization, whereby degrees of belief are strengthened rather than weakened after encountering both confirming and disconfirming evidence. Their data also support the existence of a "rebound effect," whereby participants are swayed temporarily by counter-evidence, only to revert to their former attitudes and beliefs or to even more extreme positions (Lord, Ross, & Lepper 1979, 2105). Improving on some of their methods,[13] Taber and Lodge (2006) claim to find stronger evidence for "belief polarization" when it comes to people's attitudes about highly charged political issues as gun control and affirmative action, especially when it comes to people with strong prior beliefs and those who are relatively sophisticated about the topics in question. One of the main causal factors that they identify as being responsible for this effect is that people with strong prior beliefs "evaluate supportive arguments as stronger and more compelling than opposing arguments" (Taber & Lodge 2006, 757). Similarly, Nyhan and Reifler (2010) found that members of ideological subgroups failed to revise their beliefs in the face of contrary evidence, and in some cases, strengthened their beliefs. In fact, they found a "backfire effect," whereby people maintained or strengthened their view even when confronted *only* with disconfirming evidence or arguments.[14]

[13] Taber and Lodge (2006, 756) critique the finding of belief polarization in Lord, Ross, and Lepper (1979), which is based on "subjective rather than direct measures of polarization," since they "asked subjects to report subjectively whether their attitudes had become more extreme after evaluating pro and con evidence on the efficacy of capital punishment." They claim to find evidence for belief polarization based on more objective measures.

[14] The terms "belief polarization" and "attitude polarization" tend to be used to denote strengthening or maintaining attitudes in the face of *both confirming and disconfirming* evidence. The terms "backfire effect" and "boomerang effect" tend to be used to denote strengthening or maintaining attitudes in the face of *only disconfirming* evidence. Some recent studies – for example, Wood and Porter 2019 – dispute the backfire effect, but they used simple factual statements by politicians (e.g. WMD were found in Iraq) that were then contradicted with factual corrections. Stanley, Henne,

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

They posit that this occurs because thinkers are motivated to come up with counter-arguments when they encounter disconfirming evidence, which just results in maintaining or strengthening their existing beliefs.

In sum, the tendency to maintain or strengthen beliefs when presented with evidence on both sides of an issue (or even just on the opposing side), has been widely attested in cognitive science, with a sizeable body of evidence to support it. Moreover, as we shall see in Chapter 8, some psychiatrists have posited this or a closely related cognitive disposition, "bias against disconfirmatory evidence" (BADE), in order to explain the emergence and persistence of delusions (Woodward, Moritz, Cuttler, et al. 2006). For some researchers, BADE in delusional patients is just the same tendency that exists in the general population,[15] but for others it is an accentuated or exaggerated form of a similar tendency in non-patients, and some evidence supports the view that delusional patients differ from controls in this respect (Woodward, Moritz, Cuttler, et al. 2006).[16] Hence, it may be a cognitive disposition that is present in a wide range of individuals to varying degrees. Alternatively, it may manifest itself in two varieties, one pathological and one non-pathological, with somewhat different characteristics. (This issue will be revisited in discussing psychiatric patients with delusions, specifically those diagnosed with Body Dysmorphic Disorder, in Chapter 8.)

The convergence of evidence from cognitive psychology, social psychology, political science, and psychiatry, including results obtained from a variety of experimental paradigms, suggests that a myside heuristic is widely, but perhaps not universally, manifested in the human cognitive makeup in a variety of contexts. Despite the use of different labels and taxonomic categories, a myside heuristic appears to be responsible for a variety of related effects, such as confirmation bias, disconfirmation bias, belief polarization, and the backfire effect, depending on the experimental

Yang, et al. (2020) also found no evidence of a backfire effect but they deliberately chose issues that are less contentious and emotionally charged (e.g. fracking, standardized testing) than those used in other studies (e.g. capital punishment, gun control).

[15] Maher (1988, 22) writes: "… deluded patients are like normal people – including scientists – who seem extremely resistant to giving up their preferred theories even in the face of damningly negative evidence" (cited in Woodward, Moritz, Cuttler, et al. 2006, 616).

[16] Woodward, Moritz, Cuttler, et al. (2006) devised a novel experimental test for identifying extreme cases of BADE, which involves showing participants three pictures comprising a story in reverse order, along with four possible verbal descriptions of the situation depicted. The descriptions that are most plausible given the first picture become less plausible as experimenters reveal the other two pictures, which show the same scene at earlier points in time. Delusional patients tend to stick to the initial description they selected (relative to controls), despite the disconfirming evidence.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

condition in question. The empirical evidence suggests a tentative causal model for myside heuristic, along the following lines.[17] Thinkers who have a strong prior belief backed up by evidence or arguments encounter evidence or arguments that is incongruent with those beliefs. Such thinkers have an exaggerated confidence in their own initial belief and are motivated to defend it. They make an effort to refute the incongruent evidence, devising counter-arguments, finding flaws in the reasoning, reinterpreting it in such a way that it does not contradict their belief, or otherwise coming up with reasons to dismiss it. They go on to evaluate the incongruent evidence as being weak and this causes them not to assimilate or integrate it. In turn, this leads them to maintain or strengthen their confidence in their initial belief, now that they have refuted some (possibly new) counter-arguments or contrary evidence. This tentative sketch conceives of the myside heuristic as a causal process that pertains to our reasoning or inferential capacities, with some interaction between these inferential capacities and our motivations, along the lines of "hot cognition" (Kunda 1990).[18] Even though there may be other psychological factors that lead to similar results (e.g. perceptual, attentional, or memorial mechanisms), the primary one that we have been concerned with is an inferential process that is geared to evaluating arguments for and against a particular belief. Much of the empirical evidence points to a cognitive process that leads thinkers to evaluate arguments differently based on whether they are congruent or incongruent with their own beliefs. Under a variety of conditions, experimental participants evaluate arguments confirming their beliefs differently

[17] This sketch of a causal model draws on various sources. In describing the inferential process behind the confirmation bias, Klayman (1995) mentions such aspects as: overconfidence in one's initial belief, avoidance of performing tests that are likely to contradict one's hypothesis, interpreting evidence in such a way as to favor one's own hypothesis, insufficiently revising one's confidence in one's hypothesis based on contrary evidence, and reluctance to generate novel hypotheses in the face of new evidence. Edwards and Smith (1996) identify two phenomena at play: a judgment about the strength of the evidence and an effort expended to refute it. Nickerson (1998) posits two main factors: restriction of attention to a favored hypothesis and preferential treatment of evidence supporting existing beliefs. Taber and Lodge (2006, 757) say that the bias involves evaluation of arguments, differential time and resources devoted to arguing against incongruent as opposed to congruent arguments, and a preference for searching for confirming rather than disconfirming arguments. Stanovich, West, and Toplak (2013, 259) mention that the myside bias involves the generation of evidence, evaluation of evidence, and testing of hypotheses. Hahn and Harris (2014) say that confirmation bias is an umbrella term for a variety of ways that beliefs and expectations influence the selection, retention and evaluation of evidence, which overlaps significantly with "motivated reasoning," and they link it to research on "hot" cognition.

[18] In the context of research on psychiatric delusions, Bronstein and Cannon (2017) break down the bias against disconfirming evidence (BADE) into two factors, "Evidence Integration Impairment" and "Positive Response Bias," finding that the former but not the latter is associated with delusions.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

from arguments disconfirming their beliefs, failing to revise their beliefs in the face of conflicting evidence, even strengthening their beliefs if they encounter both confirming and disconfirming evidence. There is clearly room for further research on the question of the causal network associated with the myside heuristic, particularly on the interaction between inferential and motivational processes and the possible involvement of other processes, such as perceptual, attentional, and memorial ones.[19] But we think that the cognitive process that we have sketched corresponds to a heuristic that treats evidence confirming and disconfirming one's hypotheses differentially, even though the proximal causes for such a heuristic are not fully understood.

One argument against the existence of a myside heuristic along the lines just delineated is that it would be maladaptive for humans to have such a disposition in their inferential toolkit, since it might seem to be irrational to be predisposed to treat evidence differentially depending on whether it is congruent or incongruent with one's own beliefs. The rational thing to do would surely be to treat all evidence in the same way and to follow it wherever it may point, regardless of prior beliefs. Indeed, it would seem to compromise the ability of human thinkers "to adapt effectively to changing environments" (Oswald & Grosjean 2004, 81; see also Nickerson 1998, 205–210). However, in at least some contexts and against certain background conditions, there are at least four ways in which a myside heuristic can be considered to be adaptive, in conformity with bounded or ecological rationality, or even in line with ideal theoretical norms. First, from the perspective of both ideal and bounded rationality, it is often rational to maintain one's hypothesis in the face of contrary evidence, particularly if that hypothesis has been strongly supported by past evidence and has survived other attempts at falsification or refutation (see e.g. Lord, Ross, & Lepper 1979, 2108; Nickerson 1998, 206–208; Taber & Lodge 2006, 767).[20] After all, both ideally and boundedly rational agents hold their beliefs for good reasons, so they ought not to abandon them lightly. Second, sticking to one's own beliefs in the face of countervailing evidence may lead to positive thoughts about one's judgment or opinions, generally resulting in self-affirmation (Munro & Stansbury 2009). It may be

[19] Rajsic, Wilson, and Pratt (2015) claim that there is a low-level perceptual mechanism biased toward confirmation. At some points they seem to be saying that this may reflect a general tendency toward confirmation in both perception and cognition, but at other times they indicate that there is just a similarity between the perceptual and cognitive processes.

[20] Some research finds a myside heuristic particularly or solely in sophisticated reasoners or those who hold strong opinions (e.g. Taber & Lodge 2006).

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

more comforting to cling to one's favored hypotheses, and in some cases this could have greater adaptive advantage than learning the truth about certain areas of interest (e.g. one's own abilities, the loyalty and affection of one's friends). Third, if one adopts the perspective of collective rationality and thinks of a community of thinkers along the lines of a debating society, it could be adaptive for each individual with a settled opinion to be committed strongly to that opinion, advocating for it in the face of contrary evidence, as long as a variety of hypotheses is entertained and each gets a fair shake. If such a debate is carried out for the benefit of the wider community, consisting largely of those who are not firmly convinced in any direction, and the community as a whole is allowed to decide on a course of action, this procedure may yield rational and adaptive outcomes. Something like this conception of the "marketplace of ideas" is widely thought to lead to rational decision making in the legal system and in the scientific community and is often considered to be "an efficient form of division of cognitive labor" (cf. Mercier & Sperber 2011, 65). Fourth, the adaptive advantage of a myside heuristic can also be defended if one holds that human inferential capacities have been selected for argumentation rather than reasoning (Mercier & Sperber 2011). In many situations, it may be adaptive to be persuasive, to "win friends and influence people," and persuaders who are wedded to their opinion and discount counterarguments may be more persuasive than those who are not. As Mercier and Sperber (2011, 63) put it, a confirmation bias "clearly serves the goal of convincing others." These four (not mutually exclusive) explanations of how a myside heuristic may be adaptive, and even rational, in certain contexts, shows that it might be a selected feature of our cognitive makeup rather than a dysfunction, and provides further reasons for thinking that it may be a cognitive kind. It also provides a possible etiology for the myside heuristic in terms of its distal causes.

Given that the myside heuristic can be seen to be adaptive, and indeed rational, does that mean that there is no myside *bias*, just a myside *heuristic*? Notwithstanding the arguments outlined in the previous paragraph, it is still possible that the myside heuristic may lead to systematic error in certain contexts, including some experimental contexts created in the lab. This means that there are grounds for identifying a myside bias as an offshoot of a myside heuristic. It bears emphasizing that instances of the myside heuristic that can be considered instances of a myside *bias* can only be identified against a broader background or context, including the particular task at hand, the social circumstances of the thinker, their degree of expertise, and the extent to which their own prior beliefs are justified.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

This is another instance in which a cognitive kind can only be individuated relationally with reference to a particular environmental task or social context. Distinguishing a myside bias from its underlying heuristic is only possible relative to factors external to the thinker narrowly conceived.

7.5 Conclusion

For different reasons, the search for a unifying causal role of heuristics (in general) or biases (in general) is misguided. Biases are so designated because they are a systematic deviation from a rational norm, and given the heterogeneity of rationality, there is no reason to think that various rational errors will correspond to a homogeneous cognitive kind. There is no more reason to expect that cognitive biases constitute a kind than there are grounds for thinking that all visual illusions correspond to a kind. The category of cognitive heuristics that underlie these various biases also does not seem to be unified, since there are many ways in which the cognitive system can ignore information. It might seem, then, that the heuristics and biases research program, which purports to have discovered over 100 biases (and counting), rests on the mistaken idea that *cognitive bias* and *cognitive heuristic* are kinds.

Might we find patterns within the various heuristics and biases, allowing us to identify a subset of heuristics and biases as a kind? We have examined some recent attempts to provide taxonomies of heuristics, but they do not seem, at present, to point to any kind of consensus about how the various heuristics might cluster. It will not be enough to note a family resemblance relation among some subsets of the heuristics and biases. There should be something causally relating them to one another. We have identified Stanovich's *cognitive miserliness* as a putative kind unifying a subset of heuristics, as one of the more promising subtypes. However, upon closer examination *cognitive miserliness* lacks precision, and is not, as yet, a construct that corresponds to a real cognitive kind, or so we have argued. It may be that the ways in which subsets of heuristics and biases are grouped depends crucially on where and how information is ignored within the reasoning process. A taxonomy of heuristics will depend upon cognitive architecture, and as yet, there is no agreed upon cognitive architecture of human reasoning. As such, the prospects for identifying subgroupings of heuristics and biases that correspond to cognitive kinds may be grim at the present moment. Are we suggesting that the heuristics and biases research program is misguided? No, because when it comes to specific heuristics and biases, the picture is more promising. We have argued

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press

that there are at least good grounds for positing something like a *myside heuristic* or a *myside bias*, as an inferential cognitive process involved in evaluating and judging evidence for and against one's beliefs. Similar arguments might be made for some other individual heuristics and biases, though it is unlikely that every item in the menagerie of purported cognitive heuristics and biases (e.g. the IKEA effect or the Google effect), will turn out to be genuine kinds.

A final lesson that emerges from this examination of heuristics and biases concerns the distinction between a heuristic and a bias. We started by accepting the distinction often made by cognitive scientists between a bias, which is a systematic departure from rationality, and a heuristic, which is the underlying rule or process that sometimes eventuates in a bias. When it comes to the myside bias, in particular, this means that the bias can be distinguished from the heuristic only contextually, in relation to a particular task or problem, as well as a certain history of inquiry, and other factors. That is because discounting evidence against one's own favored hypothesis can be a bias in some contexts but not in others. This means that a bias is individuated both in relation to the environment of the thinker and the thinker's etiology. Therefore, if cognitive scientists have occasion to distinguish a myside bias from a myside heuristic, the former cognitive kind is externalistically individuated, and that is the basis on which it is distinguished from the latter kind. Hence, there is no prospect of identifying it with a particular neural process or structure, at least if these are individuated in the usual way in neuroscience, without reference to the broader environment or the history of the individual. Here again, we have an instance of a good candidate for a cognitive kind that is unlikely for this reason to be reducible to a neural kind.

https://doi.org/10.1017/9781009223645.008 Published online by Cambridge University Press