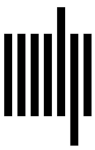
The Handbook of Rationality

Edited by: Markus Knauff, Wolfgang Spohn

Citation:

The Handbook of Rationality
Edited by: Markus Knauff, Wolfgang Spohn
DOI: 10.7551/mitpress/11252.001.0001
ISBN (electronic): 9780262366175
Publisher: The MIT Press
Published: 2021

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

1.4 Rationality and the Brain

Vinod Goel

Summary

We are widely considered the "rational animal." Coherence relations are essential for rationality. They allow for inferences that can identify actions consistent (or inconsistent) with achieving our goals, in the context of our beliefs. Consistent inferences can be further differentiated into those that are certain, plausible, and indeterminate. Twenty-plus years of research on the neural basis of reasoning reveals no single unitary mechanism of inference. Rather, our system of reasoning can be understood in terms of multiple systems for generating inferences and a common system for detecting and/or resolving conflict or inconsistency. The systems for inference generation vary as a function of conceptual and logical relations. Within logical relations, they further vary as a function of argument form, presence or absence of belief-laden content, argument presentation modality, and determinacy of the conclusion. In this chapter, we organize and present the research on the neuroscience of reasoning along these lines.

1. The Reasoning Animal

Within the Western intellectual tradition, humans are widely regarded as the reasoning or rational animal. This is to say that our behaviors are explained by postulating beliefs and desires, as well as a principle of "coherence" that guides our pursuit of the latter in the context of the former. Rationality is instrumental: it is a means to an end. A rational choice is a deliberate choice or action (selected from a large/unbounded set) that moves an organism closer to its goals in a manner consistent with its knowledge and beliefs.[1] For example, if I am thirsty and desire to drink water, and if I believe that there is a glass of water within reach on the right-hand side of my desk, and I reach out to the right-hand side, grasp it, bring it to my lips, and drink, my action is coherent or reasonable (in the context of my beliefs and desires). However, if, given the same desires and beliefs, I reach out to the left-hand side of my desk, that action would be incoherent or unreasonable because it is inconsistent with my belief that the glass is on the right-hand side and would not fulfill my desire to drink from it.[2]

I am using the term "coherence" as a basic, primitive, intuitive notion, meaning roughly "making sense." It is a relationship between propositions. Consider the following example: "If George is taller than Michael, then Michael is shorter than George." You will recognize this statement as self-evident and true. But suppose that you refuse to accept its truth and ask me to prove it. What do I do? How can I possibly prove this to you? I can't. It is like being asked to prove a postulate from Euclidean geometry. They are self-evident. Either you understand them or you don't. But we all do understand Euclidean postulates and simple logical relations as self-evident. These basic, intuitive notions are enhanced and elaborated into sophisticated systems of reasoning, involving formal logic and probability theory, that we can learn and use, with varying degrees of success.

Coherence relations can be broken down into semantic, logical, and conceptual relations. Semantic relations hold by virtue of the meaning of open-class words in the language. For example, a widow is a woman whose husband has died. Logical relations hold by virtue of the "closed-class" words in a language, such as "and," "or," "if–then," "all," "some," and "none," and prepositional phrases, such as "greater than," "inside of," and so on. Each is associated with a fixed pattern of inference. Technically, such inferences need not involve any knowledge of the world, only the language (but see below). Conceptual relations, on the other hand, involve evaluation of propositions in light of our understanding of the world, including co-occurrence experiences and causal knowledge. For example, I may conclude that all dogs have tails, because all dogs that I have seen have had tails (co-occurrence), or I may conclude that the seasons are caused by tilting of the Earth on its axis, by having a (causal) model of the Earth's orbit around the sun. We will confine our discussion to logical and conceptual relations.

Broadly speaking, the above types of inferences are our basic mechanisms for determining rational actions. They constitute a system for generating new beliefs from observations and/or existing beliefs and for maintaining consistency among our beliefs (i.e., our mental representations of the world). They allow us to generate possible actions and identify those that are consistent or inconsistent with achieving goals. Inconsistent actions can be ruled out. Consistent actions can be further broken down into those that are certain, plausible (but not certain), and indeterminate. Any creature whose actions are a function of representations of the world—in particular, representations that have propositional content—will need some system of coherence maintenance to perform these dual functions of generating inferences (to guide actions) from perceptual input and existing beliefs, as well as maintaining the consistency of beliefs.

We are creatures whose behavior is a function of our beliefs about the world rather than of the world itself. Beliefs are psychological attitudes toward mental representations that have propositional content. If I have the attitude of "belief" to the proposition "There is a tiger under my desk," then I am asserting that a certain state of affairs is true of the world (namely, that there is a tiger under my desk). The source of this knowledge can be direct perception or inference based upon perception and/or other beliefs.

Irrespective of source, to be useful and to facilitate survival and propagation, beliefs need to meet certain constraints: in the case of direct perception, they need to be veridical. In the case of inference, they need to be consistent. The issue of veridicality is largely self-evident. For instance, in the above tiger example, if my tiger beliefs are veridical with respect to the actual state of affairs, my engagement in tiger-avoiding behavior will be appropriate and conducive to my survival. If there is a mismatch between my beliefs and the facts in the world, my actions will be inappropriate. If there is no tiger under my desk but I believe there to be one, I will run away unnecessarily. If there *is* a tiger under my desk but I do not believe that there is one, I will be eaten. Beliefs that are not veridical are typically not useful (and may be harmful). Veridicality is a relationship between a representation and the world. Much perceptual and cognitive neural machinery is devoted to getting this relationship largely right, most of the time.

Apart from perception, inference constitutes the other important source of knowledge for humans. Inferences are drawn from perceptual input and/or existing belief networks. For example, suppose I observe 12 white swans on Lake Simcoe. I may be tempted to conclude, "All swans are white." This constitutes new knowledge, based on a conceptual inference from the observation of 12 swans. I then need to maintain this belief for it to guide my future behavior. There exist sophisticated long-term memory systems for this purpose. At some future date, I see a black swan at the zoo. This observation is inconsistent with my previous belief "All swans are white." If the inconsistency is detected, my belief "All swans are white" will need to be revised to "Most swans are white." In the absence of this conflict detection and revision, I would entertain the beliefs "All swans are white" and "All swans are not white" (because at least one swan is black). They cannot both be true of the world.

The importance of the consistency of beliefs is not always appreciated but is as critical as the veridicality of beliefs. For example, if I hold the belief "Tigers are extremely dangerous" and also the belief "Tigers are not extremely dangerous," what is it that I believe? More importantly, when confronted with a tiger, do I run away or do I ignore it and do nothing? Two different actions are mandated; one will lead to survival, the other to death. It is therefore not surprising that there should be considerable neural machinery devoted to maintaining consistency of beliefs.

Therefore, a creature whose actions are a function of mental representations (with propositional content) needs to ensure that these representations meet the following requirements:

1. Veridicality
2. Maintenance
3. Inference
4. Inconsistency detection
5. Updating/belief revision

The veridicality requirement falls outside the scope of inference because, as noted above, veridicality is a relationship between propositions and the world. It is largely delegated to the perceptual systems. The maintenance requirement ensures retention of beliefs and falls within the purview of memory systems. The system of inference is very much about generating new beliefs based on coherence relations between propositions. The inconsistency detection system ensures the detection of inconsistency among beliefs (which would be harmful to the organism). The belief revision requirement draws upon the inference and inconsistency detection apparatus to maintain consistency among beliefs.

The goal of this chapter is to summarize what we know about brain systems involved in inference generation and

consistency maintenance. In presenting this summary, I will not use cognitive theories of reasoning as organizing principles. Many articles and chapters organize data along one of the standard theories, be it mental models, mental logic, dual mechanisms, or a probabilistic account (Elqayam & Over, 2012; Evans, 2003; Henle, 1962; Johnson-Laird, 1994). These theories were developed to account for behavioral data. They may provide some insights for organizing the neuroscience data. However, making a priori wholesale commitments to one theory or another is probably counterproductive. My approach is to look at the neuroscience of reasoning data agnostically and to see what type of story might be embedded in it.

It is for this reason that I have tried to step back and ask, "What is the evolutionary problem that our system of inference developed to solve?" The answer that I proposed above is inference generation and inconsistency detection in service of "truth preservation." The neuroscience data on reasoning suggest that the proximal mechanisms for solving these problems are multiple systems for generating inferences and a separate common system for detecting and/or resolving inconsistency. The inference systems include a left prefrontal cortex (PFC) "interpreter" system that draws upon linguistic relations, a visuospatial system located in the parietal cortex that is engaged in linear comparisons, a visual-spatial system in the right PFC for set-inclusion determinations, and a system for dealing with indeterminate inferences, located in the right ventral lateral PFC. The generation systems do not guarantee consistency. A separate mechanism, located in the right lateral PFC, is provided for this purpose. It seems to be concerned with both the consistency between new information provided by the perceptual system and existing beliefs, as well as the consistency among propositional representations (internal or external). It is the latter issue that is relevant here.[3]

2. Neurological Systems for Generating Conceptual and Logical Inferences

The neuropsychological work on identifying systems for inference (or hypothesis generation) perhaps began with Sperry and Gazzaniga's studies on split-brain patients and Gazzaniga's conclusions regarding the dominance of the left hemisphere in generating inferences. In one classic experiment (Gazzaniga, 1989), split-brain patients were presented with a picture of a chicken claw projected to the right visual field (left hemisphere) and a picture of a snowy winter scene projected to the left visual field (right hemisphere). The patient then had to select (one with each hand), from an array of other pictures, the two most closely related to the projected pictures. The patient selected a shovel with the left hand (because the right hemisphere, controlling that hand, had processed a snowy winter scene) and a chicken with the right hand (because the left hemisphere, controlling that hand, had processed a chicken claw). Upon being asked to explain the choice of the shovel with the left hand (guided by the right hemisphere), the patient's left hemisphere (dominant for language) had no access to the information about the snowy scene processed by the right hemisphere. However, instead of responding "I don't know," the patient fabricated a plausible story, based upon background knowledge, and responded that the shovel was required to clean the chicken coop. Findings such as these led Gazzaniga (1998) to conclude that the left hemisphere was critical in drawing inferences. In fact, it couldn't help itself. It seems compelled to complete patterns and impose order on an uncertain world.

These initial findings have been fine-tuned over the years, and there is now considerable data to support the role of the left PFC in both semantic inference and simple logical inference. Unsurprisingly, however, the emerging picture is much more complex. The left PFC is not the only inference generation system in the brain. It is certainly involved in conceptual inferences. But its role in formal logical inferences seems much more constrained. Formal logical inferences engage the bilateral PFC and the parietal cortex. The relative engagement of these two regions is a function of logical form and even presentation modality. Furthermore, within the same logical form, content effects and argument determinacy modulate the specific neural systems engaged.

2.1 Conceptual Inference

Conceptual relations involve evaluation of propositions in light of our understanding of the world, including co-occurrence experiences and causal knowledge. Consider the following example:

(A) Eve is 42 years old. She is a serious and orderly woman. She loves a glass of good wine and playing chess. She tries to watch the news on foreign TV stations every day.

From this information, participants are much more likely to draw the inference "Eve is a librarian" than the inference "Eve likes to watch football." Notice that neither of these statements appear in the given information, nor follow logically from the given information. Nevertheless, the former inference is considered more

plausible than the latter, given what we have been told about Eve and what we know about the world in general (Tversky & Kahneman, 1974).

Such content-based inferences are examples of inductive inferences. They draw upon our beliefs and knowledge about the world and tend to preferentially activate the left PFC. Goel, Gold, Kapur, and Houle (1997) and Goel and Dolan (2004) carried out inductive-inference studies with arguments such as in (B) and, with respect to the PFC, reported activation in the left dorsolateral PFC, Brodmann Areas (BA) 9, 8, and 45 (figure 1.4.1a).

(B) Snakes are cold-blooded;
Alligators are cold-blooded;
∴ All reptiles are cold-blooded.

In a recent follow-up to these imaging studies, Goel, Marling, Raymont, Krueger, and Grafman (2019) had neurological patients with penetrating focal lesions engage in simple inductive inferences involving conclusions of variable believability, such as in the following arguments:

(C) Rexdale is a German shepherd;
Rexdale lives in Düsseldorf;
∴ All German shepherds live in Düsseldorf.

(D) Lipstick is moist and glossy;
Fish scales are moist and glossy;
∴ Lipstick is made from fish scales.

(E) Snakes are reptiles;
Snakes are cold-blooded;
∴ All reptiles are cold-blooded.

The conclusion of argument (C) is highly unbelievable because we know that German shepherds can be found in many cities. The conclusion of argument (E) (although technically false) is highly believable, given what we have been taught about reptiles. Most of us do not have strong beliefs about the manufacture of lipstick and thus rate the believability of the conclusion in argument (D) as less certain. The authors report that patients with unilateral focal lesions to left BA (9 and 10) have less intense beliefs about moderately believable (argument (D)) and highly believable (argument (E)) conclusions and are less likely to accept these arguments as plausible.

Inductive reasoning has also been examined through the use of analogical mapping tasks. Wharton et al. (2000) had participants examine pictures of colored geometric shapes and determine whether the shapes were analogous (analogy condition) or identical (literal condition) to a source picture of shapes. They reported enhanced brain activation in the medial frontal cortex (BA 8), the left PFC (BA 6, 10, 44, 45, 46, and 47), the anterior insula, and the left inferior parietal cortex (BA 40) when subjects made analogical-match judgments.

Other studies have imaged brain activation associated with judgment of analogous word pairs as in example (F) (Green, Fugelsang, Kraemer, Shamosh, & Dunbar, 2006) and verbal analogies, as in example (G) (Luo et al., 2003).

(F) Planet : sun versus electron : nucleus

(G) Soldier is to army as drummer is to band.

Green and colleagues (2006) found enhanced activation of parietal-frontal regions, most notably the left superior frontal gyrus (BA 9 and 10) for word pair stimuli (example (F)). Examining analogous concepts (example (G)), Luo and colleagues (2003) reported activation in the bilateral PFC (BA 45, 47, and 11), the left temporal lobe (BA 22), and the hippocampus.

Overall, studies evaluating language-based inductive arguments generally indicate activation in large areas of the brain, including the left frontal and parietal lobes. These regions overlap with the cortical regions involved in deductive reasoning with familiar material (discussed below). However, the evaluation of inductive arguments seems to be distinguished from the evaluation of deductive arguments (such as arguments (H), (N), (O), (P) below) by greater involvement of the left middle frontal gyrus (BA 9) (Goel et al., 1997; Goel & Dolan, 2004).

2.2 Logical Inference

Logical relationships between propositions are set up by closed-class (logical) terms of language rather than open-class (content) terms. For example, categorical syllogisms deal with quantification and negation, involving reasoning with the words "all," "some," and "none," as in arguments (H) and (I) below. Transitive arguments involve prepositional phrases such as "on top of," "shorter than," "more expensive than," "inside of," and "outside of" that can be used to build hierarchical relations, as in arguments (J) and (K) below. Finally, studies involving sentential connectives are generally confined to conditionals and disjunctions, as in arguments (L) and (M) below (Eimontaite et al., 2018; Noveck, Goel, & Smith, 2004). Logical arguments are designed to be valid by virtue of their structure rather than content. In terms of formal logical inference, it doesn't matter whether the arguments are about the color of broccoli or Julius Caesar crossing the Rubicon: they will be valid or invalid by virtue of their logical structure. Interestingly, this does matter to the brain. We return to this issue below.

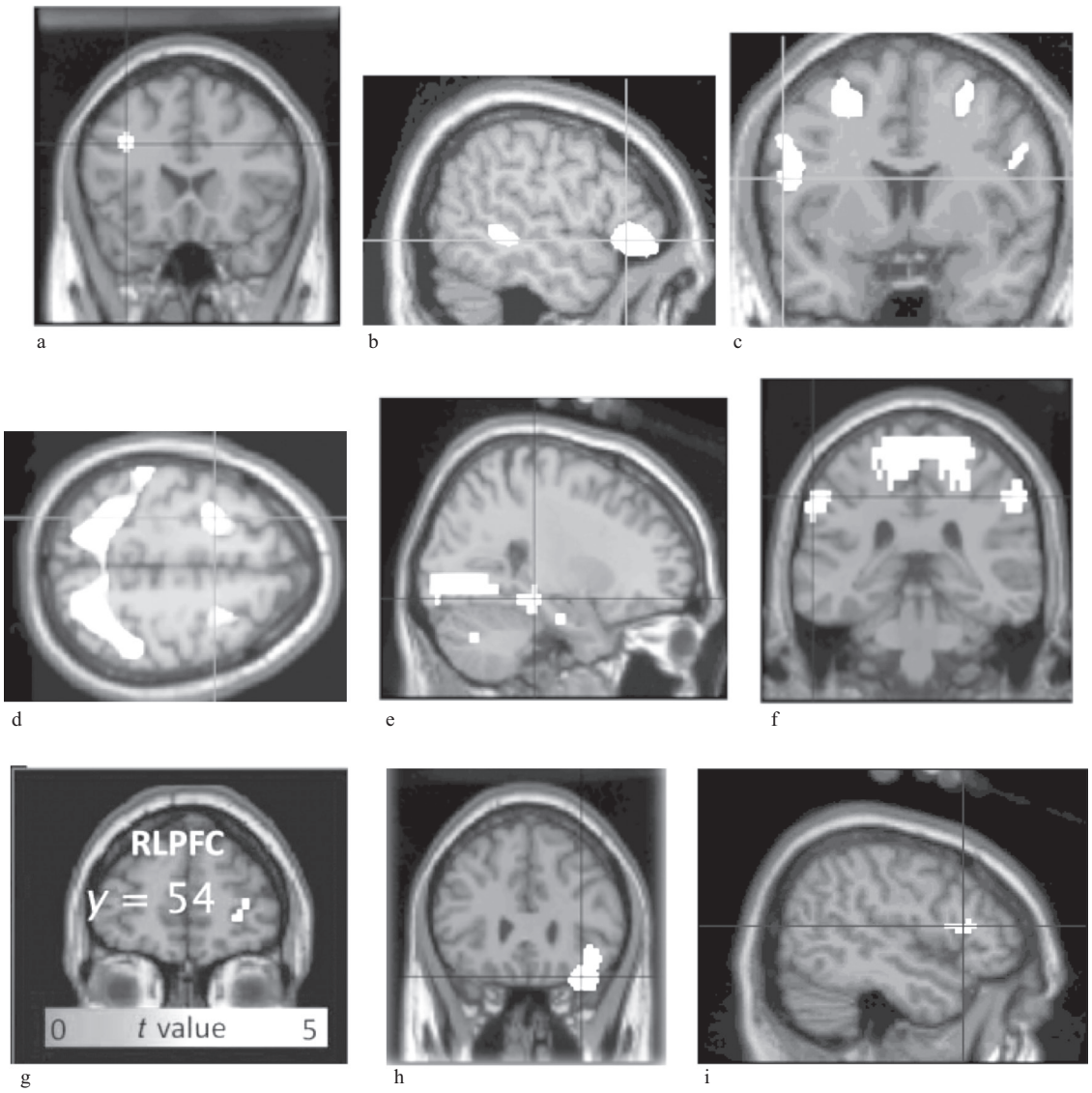


Figure 1.4.1

Systems for inference generation and conflict detection. (a–h) *Systems for inference generation*: (a) A left dorsolateral prefrontal cortex (PFC) system is involved in generating inductive inferences (reproduced with permission from Goel & Dolan, 2004); (b) a left lateral frontal-temporal linguistic system is activated during syllogistic reasoning involving content that we have beliefs about (reproduced with permission from Goel, Buchel, Frith, & Dolan, 2000); (c, d) bilateral frontal and parietal spatial systems are involved in formal syllogistic reasoning lacking meaningful semantic content (reproduced with permission from Goel et al., 2000); (e) linguistic transitive arguments with conclusions that we have beliefs about engage the left parahippocampal gyrus and the bilateral hippocampus (reproduced with permission from Goel, Makale, & Grafman, 2004); (f) linguistic transitive arguments involving conclusions that we have *no* beliefs about engage spatial systems in the bilateral parietal cortex (reproduced with permission from Goel, Makale, et al., 2004); (g) transitive arguments with pictorial stimuli engage the right rostral lateral PFC (reproduced with permission from Wendelken & Bunge, 2010); (h) indeterminate arguments with content that we have no beliefs about engage the right ventral lateral PFC (reproduced with permission from Goel, Stollstorff, Nakic, Knutson, & Grafman, 2009). (i) *A common system for conflict detection*: a common right lateral/dorsolateral PFC (BA 44/45) system seems to be engaged in detecting and/or resolving conflict or inconsistency (reproduced with permission from Goel & Dolan, 2003).

(H) All broccoli are vegetables;
All vegetables are green;
∴ All broccoli are green.

(I) All *A* are *B*;
All *B* are *C*;
∴ All *A* are *C*.

(J) London is north of Paris;
Paris is north of Rome;
∴ London is north of Rome.

(K) *A* is north of *B*;
B is north of *C*;
∴ *A* is north of *C*.

(L) If it rains on Saturday, then Linda will not come to the barbecue;
It is raining on Saturday;
∴ Linda will not come to the barbecue.

(M) Tom went to the movies with either Linda or Mary;
Tom did not go to the movies with Mary;
∴ Tom went to the movies with Linda.

Types of logical argument forms Interestingly, different logical forms seem to call upon different neural machinery. A qualitative review by Goel (2007) and a quantitative meta-analysis by Prado, Chadha, and Booth (2011) identified different brain systems engaged by categorical syllogisms, transitive arguments, and conditional arguments.

For categorical syllogisms, as in examples (H) and (I), Prado et al. (2011) report activation in the left PFC (BA 9 and 44), the left precentral gyrus (BA 4), the right caudate, and the left putamen. For arguments involving transitive relations, such as (J) and (K), they report activation in the bilateral parietal lobes (BA 7 and 40) and the bilateral dorsal PFC (BA 6). For conditional arguments, as in example (L), they report activation in a left hemisphere system involving the dorsal PFC (BA 6) and the angular gyrus (BA 39). The results regarding categorical syllogisms and transitive inference are consistent with the qualitative summary provided by Goel (2007). The results for conditional arguments may be less robust due to the insufficient number of studies. As an example, a recent study by Baggio et al. (2016) reports activation in the left PFC (BA 44 and 47) for conditional reasoning.

A number of individual neuroimaging and patient studies highlight the differential brain response to syllogistic reasoning and transitive reasoning. Activation in the left lateral and dorsal PFC (BA 6, 44, and 45) is widely reported for syllogistic reasoning tasks in neuroimaging studies (Goel, 2007; Goel et al., 2000; Goel & Dolan, 2003; Reverberi et al., 2012). One of the few lesion studies of deductive reasoning also reports that patients with left lateral and superior medial frontal lesions performed poorly on elementary deductive reasoning problems (Reverberi, Shallice, D'Agostini, Skrap, & Bonatti, 2009). Linguistically presented logical arguments involving transitive relations (examples (J) and (K)) activate the parietal cortex to a greater extent than the PFC (Goel, 2007; Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003; Modroño et al., 2018; Prado et al., 2011). In a patient study directly comparing reasoning in transitive arguments with categorical syllogisms, it was reported that patients with lesions to the parietal cortex were impaired in the former task but not the latter (Waechter, Goel, Raymont, Kruger, & Grafman, 2013).

The meta-analysis study of logical form tells us something important: there is no single mechanism for logical inference. Examination of individual studies paints an even more nuanced picture. Generation systems vary as a function of content effects, type of spatial relations involved, argument determinacy, and even modality of argument presentation (linguistic versus pictorial).

Content and logical form As indicated above, deductive arguments are valid or invalid by virtue of their logical structure. Logically, the semantic content of the premises does not make a difference to the validity of arguments. However, psychologically and neurologically, the semantic content of deductive arguments makes a significant difference. Given that we do not typically reason about *A*s and *B*s but about whether climate change causes hurricanes or whether one should buy a new or a used car, the issue of prior beliefs and semantic content of propositions becomes a central one.

Psychologically, the *content effect* is the finding that, despite deductive reasoning being a function of logical form, argument content modulates response. It is one of the oldest findings in the cognitive psychology literature. Wilkins (1928) noted that valid logical arguments with *believable* conclusions are much more likely to be rated as valid than valid arguments with *unbelievable* conclusions.

A robust consequence of the content effect is the *belief bias effect*. In reasoning with content one has beliefs about, one will encounter either a congruency or an incongruency between the logical response and conclusion believability. *Congruent* arguments are either valid with believable conclusions (argument (H)) or invalid with unbelievable conclusions (argument (N)). *Incongruent* arguments are either valid with unbelievable

conclusions (argument (O)) or invalid with believable conclusions (argument (P)).

(N) No apples are fruit;
All fruit contain calories;
∴ No apples contain calories.

(O) All apples are fruit;
All fruit are poisonous;
∴ All apples are poisonous.

(P) No apples are fruit;
All fruit contain calories;
∴ All apples contain calories.

Cognitive neuroscientists have examined content effects by comparing inferences involving meaningful contents (argument (H)) with inferences involving nonmeaningful contents (argument (I)), as well as inferences involving different types of contents (arguments (H) and (J)). These studies allow us to determine not only the effect of content on neural processing but also the effects of different *types* of contents. The neuroimaging data indicate that the main effect of comparing categorical syllogism arguments with meaningful semantic content (such as in (H)) with arguments without semantic content (such as in (I)) results in activation of a left frontal-temporal system, even after controlling for content (figure 1.4.1b) (Goel et al., 2000; Goel & Dolan, 2003). Comparing arguments without semantic content (such as (I)) with logically equivalent arguments *with* semantic content (such as (H)) activates bilateral PFC (figure 1.4.1c,d), along with bilateral parietal and occipital regions. This result persists even after controlling for the presence or absence of content (Goel et al., 2000; Goel & Dolan, 2003). When these comparisons are carried out with transitive arguments involving geographical knowledge, such as argument (J), and equivalent transitive arguments without semantic content, such as argument (K), the results show involvement of the left parahippocampal gyrus and the bilateral hippocampus in the meaningful-content condition and bilateral parietal involvement in the no-content condition, even after controlling for content (Goel, Makale, & Grafman, 2004) (figure 1.4.1e,f).

The involvement of the left frontal-temporal system in reasoning about familiar, meaningful content has also been demonstrated in neurological patients. In one study, Vartanian, Goel, Tierney, Huey, and Grafman (2009) administered three-term transitive inference arguments to patients with frontotemporal dementia. The parietal lobes of these patients were largely spared. As predicted by the imaging studies, they performed normally on arguments that they could have no beliefs about (such as (K)) but were selectively impaired in arguments involving meaningful content (such as (J)).

In another patient study, Goel, Shuren, Sheesley, and Grafman (2004) administered the Wason card selection task to patients with focal lesions to either the left or the right PFC and found that all patients performed as well as normal controls on the arbitrary version of the task, but unlike the normal controls, patients failed to benefit from the presentation of familiar content in the meaningful version of the task. In fact, consistent with the neuroimaging data, the latter result was driven by the exceptionally poor performance of patients with lesions to the left PFC. Patients with lesions to the right PFC performed as well as normal controls.

These results have been interpreted in the literature as the recruitment of different neural systems for contentful and noncontentful reasoning (Evans, 2003; Goel, 2007). However, repetitive transcranial magnetic stimulation (rTMS) studies suggest an even finer-grained distinction: Tsujii, Masuda, Akiyama, and Watanabe (2010) and Tsujii, Sakatani, Masuda, Akiyama, and Watanabe (2011) show that rTMS disruption of the left PFC specifically reduces reasoning accuracy only on a subset of contentful reasoning trials, the congruent trials. If this is the case, it suggests that the left PFC's role in logical inference may be limited to belief bias and conceptual connections and simple logical connectives. Arguments involving complex logical relations need to draw upon additional cognitive resources.

Formal logical inference Logical arguments such as (I) and (K), which lack any meaningful semantic content that participants could have beliefs about, must be evaluated with formal machinery (i.e., based purely on structure). Arguments of the forms (H), (N), (O), and (P) result in engagement of the left PFC, while logically identical arguments lacking familiar content, as in argument (I), engage the bilateral lateral PFC along with the bilateral parietal and occipital lobes (Goel et al., 2000; Goel & Dolan, 2003) (figure 1.4.1c,d).

One interpretation of these results would be that, while the left PFC may be necessary and sufficient to deal with logical inference involving familiar material that participants have beliefs about (at least in congruent trials), the right PFC is part of the system required to deal with logical inference in purely formal situations. Disruption of left PFC functioning would impair the content-sensitive inference system, resulting in poor performance on congruent trials. The rTMS data reported above support this prediction (Tsujii et al., 2010; Tsujii et al., 2011).

Stimuli presentation modality One seeming inconsistency in the neuropsychology-of-reasoning literature has to do with the neural basis of transitive inference. Above, we have reported that transitive inference involves largely the bilateral parietal cortex (Goel, 2007; Prado et al., 2011). However, at least two studies (Fangmeier, Knauff, Ruff, & Sloutsky, 2006; Wendelken & Bunge, 2010) have focused on right BA 10 (medial anterior prefrontal and right rostrolateral PFC, respectively) as critical to "relational integration" (figure 1.4.1g). Interestingly, these two studies differ from other studies of transitive reasoning in the neuroimaging literature by virtue of using nonlinguistic/pictorial stimuli. Waechter et al. (2013) propose that the modality difference across the studies (linguistic versus pictorial) may account for the differences in results. In particular, when linguistic stimuli are used, greater effort and resources are required to map the stimuli onto spatial mental models as a prerequisite to solution (Johnson-Laird, 1983; Mani & Johnson-Laird, 1982). This requires the parietal cortex (Cohen et al., 1996; Goel & Dolan, 2001; Knauff et al., 2003). In the case of pictorial stimuli, the spatial relations are actually exemplified or embodied in the stimuli; thus, this mapping has already been done in the task presentation, rendering the involvement of the parietal cortex less critical and perhaps shifting processing to the PFC.

3. Indeterminacy Tolerance

The purpose of inference is to generate new information (or at least make information that was previously implicit, explicit). For example, given the information in (Q), I can make a determinate inference about the relative population sizes of Toronto and Guelph—namely, that Toronto has a greater population than Guelph—without explicitly being told so. I can also be certain that the population of Guelph is not greater than the population of Toronto, because that would contradict the given information.

(Q) The population of Toronto is greater than the population of Hamilton. The population of Hamilton is greater than the population of Guelph.

(R) The population of Toronto is greater than the population of Hamilton. The population of Toronto is greater than the population of Guelph.

But what happens in the case of example (R)? Here we are explicitly told about the relative population sizes of Toronto and Hamilton, and of Toronto and Guelph. What inference can we draw about the relative population sizes of Guelph and Hamilton? In the absence of any additional information, relying solely on the premises, nothing follows. In this case, any conclusions we derive about the relative population sizes of Guelph and Hamilton will be invalid, not because of inconsistency with the premises but because of indeterminacy given the premises: there is no fact of the matter regarding the relative population sizes of Guelph and Hamilton, given the information provided in the premises.

Goel et al. (2007) show that neurological patients with focal lesions to the left PFC have generalized impaired reasoning, including arguments with complete information (i.e., determinate, as in argument (Q)), while patients with right PFC lesions are selectively impaired only in arguments with incomplete information (i.e., indeterminate, as in argument (R)). This patient study demonstrates a double dissociation across left and right PFC along the dimension of determinacy. Neuroimaging studies involving similar transitive arguments reveal similar results (figure 1.4.1h) (Brzezicka et al., 2011; Goel et al., 2009). A study examining deductive reasoning with indeterminate syllogistic arguments also reported activation in the right PFC instead of the left PFC (Parsons & Osherson, 2001). (A transcranial magnetic stimulation study involving spatial relational reasoning also showed involvement of the right superior parietal cortex in dealing with uncertainty, although it did not test for the involvement of the right PFC; Ragni, Franzmeier, Maier, & Knauff, 2016.)

These data seem to suggest that we have developed special brain systems for dealing with indeterminate inferences. In cases where we can, we will fill in the missing information, using the left hemisphere interpreter, but in cases where this is not possible, a right ventral-lateral PFC system is engaged to tolerate or accommodate the indeterminacy.

4. Detection of Conflict or Inconsistency

Conflict or inconsistency can arise among one's existing beliefs, between existing beliefs and incoming new information, and from inferences drawn from existing beliefs or external propositions. Detecting and/or resolving these conflicts or inconsistencies seems to be a generalized function of the right lateral and dorsolateral PFC (figure 1.4.1i).[4]

In the context of logical reasoning, we are specifically referring to an inconsistency between a response cued by our beliefs about the world and a response cued by the logical structure of the argument. Within incongruent trials (as in arguments (O) and (P)), the prepotent response is the incorrect response associated with the

believability of the conclusion. Incorrect responses in such trials indicate that subjects failed to detect and/or overcome the conflict between their beliefs and the logical inference and/or to inhibit the prepotent response associated with the belief bias. These belief-biased responses activate the ventral-medial PFC (BA 11 and 32), highlighting its role in nonlogical, belief-based responses (Goel & Dolan, 2003). The correct response indicates that subjects detected the conflict between their beliefs and the logical inference, inhibited the prepotent response associated with the belief bias, and engaged a formal reasoning mechanism. The detection of this conflict requires engagement of the right lateral and the dorsal-lateral PFC (BA 45 and 46) (see figure 1.4.1i), while generating the logical response calls upon the visuospatial machinery in the parietal cortex (De Neys, Vartanian, & Goel, 2008; Goel et al., 2000; Goel & Dolan, 2003; Prado & Noveck, 2007; Stollstorff, Vartanian, & Goel, 2012). Knauff and colleagues (Hamburger et al., 2018; Knauff, 2013) make a related point in the context of visual imagery impeding spatial reasoning.

These functional magnetic resonance imaging results have been replicated by rTMS studies demonstrating that stimulation of the right PFC specifically impairs performance on incongruent reasoning trials such as in arguments (O) and (P) (Tsujii et al., 2010; Tsujii et al., 2011). The rTMS data also show that disruption of the left PFC results not only in decreased performance in congruent trials but also in *improved* performance in *in*congruent trials. That is, when the left PFC is impaired, participants are less likely to go with the believability of the conclusion and will recruit other cortical regions to formally evaluate the argument.

One early demonstration of this conflict detection system with lesion data was carried out by Caramazza, Gordon, Zurif, and DeLuca (1976) using simple two-term reasoning problems such as the following: "Mike is taller than George. Who is taller? Who is shorter?" They reported that left hemisphere patients were impaired in all forms of the problem but—consistent with imaging data (Goel et al., 2000; Goel & Dolan, 2003; Stollstorff et al., 2012)—right hemisphere patients were only impaired when the form of the question was inconsistent with the premise (i.e., "Who is shorter?").

A final example of the role of the right PFC in conflict detection is provided by Reverberi, Lavaroni, Gigli, Skrap, and Bonatti (2005). They carried out a revised version of the Brixton task involving rule-induction and rule-conflict conditions. They reported that while patients with lesions to the left PFC showed an impairment in rule induction, patients with lesions to the right PFC were impaired specifically in the rule-conflict condition.

This conflict detection role of the right lateral and the dorsal PFC is a generalized phenomenon that has been documented in a wide range of paradigms in the cognitive neuroscience literature (Fink et al., 1999; Picton, Stuss, Shallice, Alexander, & Gillingham, 2006; Vallesi, Mussoni et al., 2007; Vallesi, Shallice, & Walsh, 2007). Marinsek, Turner, Gazzaniga, and Miller (2014) have actually suggested that conflict or inconsistency detection is the main role of the right PFC in cognitive functioning.

5. Summary and Conclusion

Humans are creatures whose behavior is a function of their beliefs about the world rather than the world itself. Not only do we have beliefs, but our beliefs also have propositional content. Beliefs with propositional content allow us to generate new knowledge by drawing inferences that take us beyond direct perception and differentiate between what is necessarily the case, what might be the case, and what absolutely *cannot* be the case. One can even time-travel with such mental representations and entertain past and future possibilities, including counterfactuals.

For such a system to be useful, the inferences must be coherent, and this coherence must be maintained over the whole system of beliefs. There must also be a system for detecting inconsistencies in inferences and among beliefs. Twenty-plus years of neuroscience studies of logical reasoning have revealed that there is no unitary reasoning module in the brain for undertaking this. Rather, our ability to reason seems to be underwritten by two separate classes of mechanisms: (1) mechanisms for hypothesis generation and inference and (2) a mechanism for detecting conflict or inconsistency.

Inference generation calls upon several different systems, including (1) a left PFC interpreter system sensitive to semantic, conceptual, and simple logical relations; (2) multiple visuospatial systems; and (3) a system for tolerating indeterminacy. The first of these systems deals largely with semantic and conceptual relations and simple syntactic inferences. The second deals with more formal processing of logical arguments. The neurological basis of this system seems to vary as a function of logical form and argument presentation modality. For example, set-inclusion relationships call upon the right and/or bilateral PFC, while linear comparisons call upon the parietal cortex. Additionally, linear comparisons presented pictorially activate the right rostral PFC, while the same linear comparisons presented linguistically

activate bilateral parietal cortex. Finally, a system in the right ventral-lateral PFC seems to play a critical role in allowing for indeterminate inferences.

These systems of inference generation do not, in and of themselves, guarantee consistency. We seem to have a separate system in the right lateral/dorsolateral PFC for detecting conflict or inconsistency between external and internal representations and among internal representations. Together, these various systems account for the ability to draw correct inferences and maintain the consistency of the overall belief network.

The details of this overall account will undoubtedly change as additional studies are carried out and new data are generated and added to our knowledge base. However, after 20 years of research, the broader picture that is emerging may be reasonably secure: specifically, that there is no single system of logical reasoning in the brain. Our ability to engage in rational thought is underwritten by several different types of inference systems and a common system for detecting inconsistencies.

Notes

1. Additionally, while not germane to our purposes here, it is important to note that there is a "gap" between input and output conditions in rational actions. The antecedent condition is never causally sufficient for the consequent condition (Cassirer, 1944).

2. Sometimes a distinction is made between "instrumental" rationality and "epistemic" rationality (Stanovich, 1999). The latter is an evaluation of the fit between an individual's beliefs and the facts in the world. However, this is simply instrumental rationality applied to belief evaluation and revision. It is not clear that this distinction needs to be dealt with separately, at least for our purposes.

3. Most neuroscience discussions of rationality usually begin and end with delusions. While delusions are an important topic, I will not touch upon them here. I think we will increase our chances of understanding delusions if we can first understand the neural basis of normal rationality.

4. Experimental data have thus far not clearly distinguished between the right PFC's role in detecting and resolving the conflict.

References

Baggio, G., Cherubini, P., Pischedda, D., Blumenthal, A., Haynes, J.-D., & Reverberi, C. (2016). Multiple neural representations of elementary logical connectives. *NeuroImage*, *135*, 300–310.

Brzezicka, A., Sędek, G., Marchewka, A., Gola, M., Jednoróg, K., Królicki, L., & Wróbel, A. (2011). A role for the right prefrontal and bilateral parietal cortex in four-term transitive reasoning: An fMRI study with abstract linear syllogism tasks. *Acta Neurobiologiae Experimentalis*, *71*, 479–495.

Caramazza, A., Gordon, J., Zurif, E. B., & DeLuca, D. (1976). Right-hemispheric damage and verbal problem solving behavior. *Brain and Language*, *3*, 41–46.

Cassirer, E. (1944). *An essay on man: An introduction to a philosophy of human culture*. New Haven, CT: Yale University Press.

Cohen, M. S., Kosslyn, S. M., Breiter, H. C., DiGirolamo, G. J., Thompson, W. L., Anderson, A. K., . . . Belliveau, J. W. (1996). Changes in cortical activity during mental rotation: A mapping study using functional MRI. *Brain*, *119*, 89–100.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*, *19*, 483–489.

Eimontaite, I., Goel, V., Raymont, V., Krueger, F., Schindler, I., & Grafman, J. (2018). Differential roles of polar orbital prefrontal cortex and parietal lobes in logical reasoning with neutral and negative emotional content. *Neuropsychologia*, *119*, 320–329.

Elqayam, S., & Over, D. (2012). Probabilities, beliefs, and dual processing: The paradigm shift in the psychology of reasoning. *Mind & Society*, *11*, 27–40.

Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Science*, *7*, 454–459.

Fangmeier, T., Knauff, M., Ruff, C. C., & Sloutsky, V. (2006). fMRI evidence for a three-stage model of deductive reasoning. *Journal of Cognitive Neuroscience*, *18*, 320–334.

Fink, G. R., Marshall, J. C., Halligan, P. W., Frith, C. D., Driver, J., Frackowiak, R. S. J., & Dolan, R. J. (1999). The neural consequences of conflict between intention and the senses. *Brain*, *122*, 497–512.

Gazzaniga, M. S. (1989). Organization of the human brain. *Science*, *245*, 947–952.

Gazzaniga, M. S. (1998). *The mind's past*. Berkeley: University of California Press.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Science*, *11*, 435–441.

Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, *12*, 504–514.

Goel, V., & Dolan, R. J. (2001). Functional neuroanatomy of three-term relational reasoning. *Neuropsychologia*, *39*, 901–909.

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition*, *87*, B11–B22.

Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition*, *93*, B109–B121.

Goel, V., Gold, B., Kapur, S., & Houle, S. (1997). The seats of reason? A localization study of deductive and inductive reasoning. *NeuroReport*, *8*, 1305–1310.

Goel, V., Makale, M., & Grafman, J. (2004). The hippocampal system mediates logical reasoning about familiar spatial environments. *Journal of Cognitive Neuroscience*, *16*, 654–664.

Goel, V., Marling, M., Raymont, V., Krueger, F., & Grafman, J. (2019). Patients with lesions to left prefrontal cortex (BA 9 and BA 10) have less entrenched beliefs and are more sceptical reasoners. *Journal of Cognitive Neuroscience*, *31*, 1674–1688.

Goel, V., Shuren, J., Sheesley, L., & Grafman, J. (2004). Asymmetrical involvement of frontal lobes in social reasoning. *Brain*, *127*, 783–790.

Goel, V., Stollstorff, M., Nakic, M., Knutson, K., & Grafman, J. (2009). A role for right ventrolateral prefrontal cortex in reasoning about indeterminate relations. *Neuropsychologia*, *47*, 2790–2797.

Goel, V., Tierney, M., Sheesley, L., Bartolo, A., Vartanian, O., & Grafman, J. (2007). Hemispheric specialization in human prefrontal cortex for resolving certain and uncertain inferences. *Cerebral Cortex*, *17*, 2245–2250.

Green, A. E., Fugelsang, J. A., Kraemer, D. J. M., Shamosh, N. A., & Dunbar, K. N. (2006). Frontopolar cortex mediates abstract integration in analogy. *Brain Research*, *109*, 125–137.

Hamburger, K., Ragni, M., Karimpur, H., Franzmeier, I., Wedell, F., & Knauff, M. (2018). TMS applied to V1 can facilitate reasoning. *Experimental Brain Research*, *236*, 2277–2286.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review*, *69*, 366–378.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1994). Mental models, deductive reasoning, and the brain. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 999–1008). Cambridge, MA: MIT Press.

Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. Cambridge, MA: MIT Press.

Knauff, M., Fangmeier, T., Ruff, C. C., & Johnson-Laird, P. N. (2003). Reasoning, models, and images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, *15*, 559–573.

Luo, Q., Perry, C., Peng, D., Jin, Z., Xu, D., Ding, G., & Xu, S. (2003). The neural substrate of analogical reasoning: An fMRI study. *Cognitive Brain Research*, *17*, 527–534.

Mani, K., & Johnson-Laird, P. N. (1982). The mental representation of spatial descriptions. *Memory & Cognition*, *10*, 181–187.

Marinsek, N., Turner, B. O., Gazzaniga, M., & Miller, M. B. (2014). Divergent hemispheric reasoning strategies: Reducing uncertainty versus resolving inconsistency. *Frontiers in Human Neuroscience*, *8*, 839.

Modroño, C., Navarrete, G., Nicolle, A., González-Mora, J. L., Smith, K. W., Marling, M., & Goel, V. (2018). Developmental grey matter changes in superior parietal cortex accompany improved transitive reasoning. *Thinking & Reasoning*, *25*, 151–170.

Noveck, I. A., Goel, V., & Smith, K. W. (2004). The neural basis of conditional reasoning with arbitrary content. *Cortex*, *40*, 613–622.

Parsons, L. M., & Osherson, D. (2001). New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex*, *11*, 954–965.

Picton, T. W., Stuss, D. T., Shallice, T., Alexander, M. P., & Gillingham, S. (2006). Keeping time: Effects of focal frontal lesions. *Neuropsychologia*, *44*, 1195–1209.

Prado, J., Chadha, A., & Booth, J. R. (2011). The brain network for deductive reasoning: A quantitative meta-analysis of 28 neuroimaging studies. *Journal of Cognitive Neuroscience*, *23*, 3483–3497.

Prado, J., & Noveck, I. A. (2007). Overcoming perceptual features in logical reasoning: A parametric functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, *19*, 642–657.

Ragni, M., Franzmeier, I., Maier, S., & Knauff, M. (2016). Uncertain relational reasoning in the parietal cortex. *Brain and Cognition*, *104*, 72–81.

Reverberi, C., Bonatti, L. L., Frackowiak, R. S. J., Paulesu, E., Cherubini, P., & Macaluso, E. (2012). Large scale brain activations predict reasoning profiles. *NeuroImage*, *59*, 1752–1764.

Reverberi, C., Lavaroni, A., Gigli, G. L., Skrap, M., & Shallice, T. (2005). Specific impairments of rule induction in different frontal lobe subgroups. *Neuropsychologia*, *43*(3), 460–472.

Reverberi, C., Shallice, T., D'Agostini, S., Skrap, M., & Bonatti, L. L. (2009). Cortical bases of elementary deductive reasoning: Inference, memory, and metadeduction. *Neuropsychologia*, *47*, 1107–1116.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. New York, NY: Psychology Press.

Stollstorff, M., Vartanian, O., & Goel, V. (2012). Levels of conflict in reasoning modulate right lateral prefrontal cortex. *Brain Research*, *1428*, 24–32.

Tsujii, T., Masuda, S., Akiyama, T., & Watanabe, S. (2010). The role of inferior frontal cortex in belief-bias reasoning: An rTMS study. *Neuropsychologia*, *48*, 2005–2008.

Tsujii, T., Sakatani, K., Masuda, S., Akiyama, T., & Watanabe, S. (2011). Evaluating the roles of the inferior frontal gyrus and superior parietal lobule in deductive reasoning: An rTMS study. *NeuroImage*, *58*, 640–646.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Vallesi, A., Mussoni, A., Mondani, M., Budai, R., Skrap, M., & Shallice, T. (2007). The neural basis of temporal preparation: Insights from brain tumor patients. *Neuropsychologia*, *45*, 2755–2763.

Vallesi, A., Shallice, T., & Walsh, V. (2007). Role of the prefrontal cortex in the foreperiod effect: TMS evidence for dual mechanisms in temporal preparation. *Cerebral Cortex*, *17*, 466–474.

Vartanian, O., Goel, V., Tierney, M., Huey, E. D., & Grafman, J. (2009). Frontotemporal dementia selectively impairs transitive reasoning about familiar spatial environments. *Neuropsychology*, *23*, 619–626.

Waechter, R. L., Goel, V., Raymont, V., Kruger, F., & Grafman, J. (2013). Transitive inference reasoning is impaired by focal lesions in parietal cortex rather than rostrolateral prefrontal cortex. *Neuropsychologia*, *51*, 464–471.

Wendelken, C., & Bunge, S. A. (2010). Transitive inference: Distinct contributions of rostrolateral prefrontal cortex and the hippocampus. *Journal of Cognitive Neuroscience*, *22*, 837–847.

Wharton, C. M., Grafman, J., Flitman, S. S., Hansen, E. K., Brauner, J., Marks, A., & Honda, M. (2000). Toward neuroanatomical models of analogy: A positron emission tomography study of analogical mapping. *Cognitive Psychology*, *40*, 173–197.

Wilkins, M. C. (1928). *The effect of changed material on the ability to do formal syllogistic reasoning* (Archives of Psychology No. 102). New York, NY: Woodworth.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Knauff, Markus, editor. | Spohn, Wolfgang, editor.
Title: The handbook of rationality / edited by Markus Knauff and Wolfgang Spohn.
Description: Cambridge : The MIT Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020048455 | ISBN 9780262045070 (hardcover)
Subjects: LCSH: Reasoning (Psychology) | Reason. | Cognitive psychology. | Logic. | Philosophy of mind.
Classification: LCC BF442 .H36 2021 | DDC 153.4/3—dc23
LC record available at https://lccn.loc.gov/2020048455