This is a section of doi:10.7551/mitpress/11252.001.0001

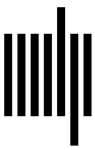
The Handbook of Rationality

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

2.3 Mental Models, Reasoning, and Rationality

P. N. Johnson-Laird

Summary

Is human reasoning rational? This chapter argues that it is not based on logic or the probability calculus. Instead, rational inferences aim for accurate models of the world. An *intuitive* system uses mental models that ignore what is false in possibilities. So, it succumbs to systematic fallacies. But a *deliberative* system relies on explicit models that also represent what is false in possibilities, and so it can correct the fallacies. Neither system knows how to infer the probability of compound assertions, such as conjunctions. So, inferences often violate the probability calculus. In tests of inductive hypotheses, intuitions aim to establish examples of them; with feedback from tests, deliberations aim also to show no counterexamples exist. Reasoners can abduce algorithms and explanations that resolve inconsistencies between facts and inferences. The biggest risk in all reasoning is to overlook a possibility, but a simple procedure helps to prevent this sort of irrationality.

> Science can never grapple with the irrational.
> —Oscar Wilde, *An Ideal Husband*

1. Rationality and Why It Is Not Based on Logic

Biden will impose new tariffs on Russia or China, or both of them.

So, it is possible that he will impose new tariffs on Russia.

People make this inference, and it seems sensible. But, as you will soon see, it violates any logic dealing with possibilities. The inference raises the questions of what counts as rational in reasoning and to what extent naïve individuals—those untrained in logic or cognate disciplines—meet this target.

Certain beliefs, inferences, and actions are irrational. For instance, it is irrational to believe that the Earth is flat (Sperber, 1995), to infer that the bow doors of a car ferry are closed because no one has reported otherwise (Yardley, 2014), or, like the man who kissed his pet, to fail to think of an obvious possibility (his pet was a rattlesnake; see Northcutt, 2002). Yet, theorists disagree about what counts as rational thinking (see section 1 of this volume). It could be, for example, any mental process that is

- conducive to the survival of your genes and those of your relatives,
- predisposed to yield truth, and
- based on logic, the probability calculus, decision theory, or game theory.

The unqualified use of "logic" here and throughout this chapter refers to classical logic (see chapter 3.1 by Steinberger, this handbook).

It is hard to resolve the theorists' disagreement, because the only satisfactory way to do so should itself be rational. And, alas, that is what's at stake. This difficulty has sometimes led psychologists to describe inferences rather than to evaluate them (cf. chapter 1.2 by Evans, in this handbook). But if psychologists want to improve reasoning, they cannot ignore criteria for rationality. So, what are they?

Once upon a time, the answer was simple. It was: reasoning based on the laws of thought or, as Boole (1854) remarked, the laws of "right reasoning," which he formalized in a single algebra for both logic and probability. It should yield *valid* inferences, which are those whose conclusions are true in every case in which all their premises are true (Jeffrey, 1981, p. 1). It also requires that there is at least one case in which all the premises are true. Otherwise, contradictions imply any conclusion whatsoever, and naïve individuals do not make such inferences. The preceding definition of validity is independent of logic. It hinges on the meanings of assertions, as the following inference illustrates:

Necessarily, the plants will die.

Therefore, the plants will die.

It is valid in one logic of possibilities, but not in another, because the meaning of "necessarily" differs between the two logics (Girle, 2009).

Logic depends on formal rules that yield proofs of valid inferences, and psychologists once took for granted that logic underlies everyday reasoning (see chapter 3.2 by O'Brien, this handbook). But an immediate problem is that logic concerns the relations between sentences in a formal language, whereas inferences in daily life concern the propositions that utterances in natural language convey in context. To apply a formal rule of inference, such as *A or B or both*; *not A*; therefore, *B*, to propositions, you need to know their "logical forms." No algorithm yet exists that can carry out this task; it is extraordinarily difficult (see Keene, 1992). Consider, for example, the disjunctive premise in our opening example: *Biden will impose new tariffs on Russia or China, or both of them*. You understand that "or China" abbreviates the clause, "Biden will impose new tariffs on China." So, the logical form of the premise matches *A or B or both* in the formal rule above. This understanding depends on knowledge. And once you take that into account, there is no need for logical form. As we'll see, reasoning can be based on representations of the meaning and reference of assertions. Another difficulty is that sensible inferences in daily life often diverge from logic. The opening example calls for a logic dealing with possibilities, of which there are infinitely many (Girle, 2009). Yet, the inference is invalid in all of them (Johnson-Laird & Ragni, 2019). In logic, the disjunction is true given that at least one of its two clauses is true. But it isn't valid to infer just one clause from a disjunction of two. Imagine, say, that it is impossible that Biden will impose new tariffs on Russia but true that he will impose them on China. In this case, the disjunctive premise is true, but the conclusion is false, and so the inference is invalid in logic. We see later why people nonetheless make it and why some psychologists have given up logic in favor of the probability calculus as the basis of reasoning (cf. section 4 of this volume).

The present chapter advocates neither logic nor probability as the foundation of reasoning. Instead, its starting point is evolution. Living creatures evolved with a perceptual apparatus that constructs internal models of the world. The word "model" here signifies representations that are finite and iconic in that their structure mirrors the structure of the world. These models were adaptive for our evolutionary ancestors, and we have inherited the ability to construct them. And in almost any situation, we have the ability to conceive of a small exhaustive set of possibilities. Anything common to all of them is certain, unless they concern what is permissible—in which case, it is obligatory. Craik (1943) wrote that our minds construct small-scale models of the world to help us to make decisions but that reasoning depends on verbal rules. The modern "model" theory assumes that reasoning relies instead on iconic models, which represent possibilities (Johnson-Laird, 1983). It postulates this working definition of right reasoning:

Rational inferences yield accurate models both of possibilities and of their implications.

The chapter overlooks strategic reasoning in games (see section 9 of this volume); as the late Reinhard Selten once remarked, "Game theory is rational theology" (Steingold & Johnson-Laird, 2002). Otherwise, it deals with all the main sorts of reasoning: deductive, probabilistic, inductive, and abductive.

2. Deduction and Possibilities

2.1 The Theory of Mental Models

The theory postulates a mechanism that builds models from the premises of inferences, which can be perceptions, descriptions, knowledge, or beliefs. So, a premise such as

There's a circle or a triangle, or both, in the picture

has a set of three models of the possibilities in the picture:

○
Δ
○ Δ

Together, they have the force of an exhaustive conjunction of three possibilities that each hold in default of knowledge to the contrary. And so the first model yields the inference

Therefore it is possible that there is a circle in the picture.

The inference is valid by default, and individuals make it (Hinterecker, Knauff, & Johnson-Laird, 2016). They also estimate the conclusion as less probable than that of the premise, which is contrary to probabilistic approaches to reasoning (see below). Defaults are commonplace. You infer that a bird flies until you discover that it is an emu. The model theory extends the concept to the meaning of connectives, such as "or."

Logic stipulates that a conditional, *If A then B*, is true in every case except when *A* is true and *B* is false—in this case alone, the conditional is false. The general claim

If anything is a swan then it is white

is thus equivalent in logic to

If anything isn't white then it isn't a swan.

So, a green carnation is corroboratory evidence for the claim: it isn't white, and it isn't a swan. This well-known "paradox" of confirmation holds for logic (Hempel, 1945).

In contrast, the model theory treats the conditional about swans as true provided that examples are possible (swans that are white), but counterexamples are impossible (swans that are not white). Things that are not swans are possible whether the conditional is true or false (Johnson-Laird, Khemlani, & Goodwin, 2015). So, carnations are irrelevant. Likewise, scientists need examples of a hypothesis that they are trying to induce or test from observations (Nicod, 1950). Science, as many major practitioners have acknowledged, is a refinement of everyday thinking (see chapter 14.1 by Andersen & Andersen, this handbook). And a famous test illustrates scientific practice. The general theory of relativity predicts that if a ray of light passes close to a heavenly body, it should deviate toward the latter by a precise amount depending on its distance from the heavenly body's center (Einstein, 1920/2004, p. 111). To test this hypothesis, astronomers measured the apparent positions of "fixed" stars when light from them passed close to the sun—a situation satisfying the *if*-clause of the hypothesis—to determine whether examples of the hypothesis occurred and counterexamples did not. The results corroborated general relativity.

Reasoning based on models can deliver valid inferences, and validity is part of right reasoning. But it is not enough. The validity of the following inference is undeniable:

It's raining.

Therefore, it's raining and it's raining.

Yet, it is useless, except as an illustration of logic. In rational inferences, individuals should draw conclusions that are parsimonious (not just conjunctions of premises), that are novel (not just a repetition of a premise), and that maintain the semantic information in the premises, which guarantees validity (Johnson-Laird & Byrne, 1991). These are all emergent properties of inferences based on models. One notable consequence occurs with premises such as

None of the teachers is poor.

None of the parents is poor.

Most people respond, "Nothing follows." The premises establish no definite relation that meets the constraints above (see chapter 3.3 by Khemlani, this handbook). But the response is contrary to logic, which allows that infinitely many valid conclusions follow from any set of premises. They include some that people might accept if they were presented to them (e.g., "It is possible that some of the teachers are parents"). And many deductions are downright stupid. The model theory distinguishes those that reasoners draw for themselves within those that are valid.

2.2 Irrational Deductions

Do naïve individuals reason rationally? Even Boole (1854) recognized that people err, and so the nature of their mistakes matters. Haphazard errors don't threaten rationality. They fail to reflect underlying competence. But the model theory predicts systematic fallacies. It assumes that intuitive inferences rely on a simpler system of reasoning than deliberations: it postulates two systems of reasoning, which can lead to different conclusions. Wason was the first to formulate such a dual-system theory, but unlike other theories (see chapter 1.2 by Evans and chapter 2.5 by Klauer, both in this handbook), both its intuitive and deliberative systems were described in an algorithm (Johnson-Laird & Wason, 1970a).

The crux for rationality is whether reasoners rely on *mental* models or on *explicit* models. As earlier, the set of mental models for the disjunction "There's a circle or a triangle, or both" is:

```
o
    Δ
o   Δ
```

The first model represents the possibility that there is a circle but nothing about triangles. Mental models in the intuitive system represent only what is true in each possibility. In contrast, explicit models in the deliberative system use negation to represent clauses in the premises that are false in a possibility. The explicit models of the preceding disjunction are therefore:

```
 o  ¬Δ
¬o   Δ
 o   Δ
```

where "¬" denotes the mental symbol for negation, which is linked to its semantics.

The fallacies from mental models are ubiquitous (Khemlani & Johnson-Laird, 2017). Here is an example from two *exclusive* disjunctions, that is, those in which not both of their clauses can hold:

Either the circle is in the picture or else the triangle is.

Either the circle is in the picture or else the triangle isn't.

Could both of these assertions be true at the same time? Most people say, "Yes." They rely on mental models and grasp that the circle is possible according to each premise. They overlook that when one clause of an exclusive disjunction is true, the other clause is false. So, the explicit models of the two premises tell a different story. The first premise has the explicit models

```
 o  ¬Δ
¬o   Δ
```

And the second premise has the explicit models

o Δ
¬o ¬Δ

The two sets have no possibility in common. And so the two disjunctions cannot both be true. Mental models lead to predictable fallacies, as in this example, but explicit models can correct them—in principle, they yield only deductions that are valid according to the model theory. They demand more capacity in working memory, and an inference can exceed that capacity—a factor that differs from one person to another. Deductions based on "or" and other connectives are computationally intractable—the required processing for many distinct premises can exceed the capacity of a finite brain, even one as big as the universe (see the Overview by Knauff & Spohn, this handbook).

The model theory applies to all the main domains of deductive reasoning (see chapter 3.3 by Khemlani on syllogisms, chapter 6.3 by Byrne & Espino on counterfactual reasoning, chapter 13.2 by Ragni on spatial and temporal relations, and chapter 13.3 by Knauff on visualization, all in this handbook). And its intuitive and deliberative processes have been implemented in a computer program for reasoning with connectives (mSentential, available at https://mentalmodels.princeton.edu/models/). It takes as input premises in natural language and evaluates the validity of a putative conclusion as necessary, possible, or impossible. From simple premises, it can construct its own conclusion in keeping with the constraints on rationality described earlier. It embodies the idea that individuals can make valid deductions, but in certain cases, they succumb to illusory inferences. They are a consequence of mental models, which represent only what is true in possibilities. This focus may reflect the perceptual origins of models.

3. Probability

3.1 Probabilistic Theories of Reasoning

The probability calculus is nothing more than common sense according to Laplace (1951/1820). This view is embodied in psychological theories that replace logic with probabilities as the foundation of everyday reasoning (see chapter 4.4 by Pfeifer, chapter 4.5 by Chater & Oaksford, and chapter 4.6 by Oberauer & Pessach, all in this handbook). On this "probabilist" account, the inference

If it rained then it was cold

It rained

Therefore, it was cold

depends neither on logical rules nor on mental models but on probabilities. These theories hark back to a treatment of probabilistic validity (p-validity) applying to conditionals (Adams, 1998). In its simplest definition, an inference is p-valid if its conclusion is no less probable than its premises in any consistent assignment of probabilities to its clauses.

Probabilists have criticized the model theory for diverging from orthodox logic (e.g., Oaksford, Over, & Cruz, 2019). But, by design, the model theory differs from logic. And probabilism has difficulties of its own, which illuminate the nature of rationality. One difficulty is that possibilities can underlie probabilities, but probabilities cannot underlie possibilities. Consider a museum attendant who gives this permission to a visitor:

If you have a ticket you can enter now or in fifteen minutes.

Knowing that she has a ticket, the visitor infers:

I can enter now.

The inference is sensible, and the model theory predicts it. Yet, it violates logic to infer one clause from a disjunction, and the inference is also p-invalid. But, as probabilists can point out, probabilities don't apply to permissions. However, the same sort of inference occurs with descriptions:

If the weather is bad, then it's possible that there's flooding or a tornado.

The weather is bad.

So, it's possible that there's flooding.

The model theory assigns probabilities to possibilities, and other methods exist to do so too (e.g., Lewis, 1981). And the preceding inference creates a dilemma for probabilism: either a probabilistic account does not apply to it, or else it violates p-validity.

How do naïve individuals infer probabilities? The answer is: with difficulty. And probabilist theories have so far failed to explain the process. Consider two assertions:

A: Trump is the Republican presidential candidate in 2024.

B: Warren is the Democratic presidential candidate in 2024.

Given such assertions, participants are happy to make estimates of the following triples of probabilities (Khemlani, Lotstein, & Johnson-Laird, 2015):

- P(*A*), P(*B*), and P(*A* and *B*).
- P(*A*), P(*B*), and P(*A* or *B* or both).
- P(*A*), P(*B*), and P(*B* given *A*).

Each set fixes the complete joint probability distribution (JPD) for *A* and *B* i.e., the probabilities of each of these four exhaustive cases: *A* & *B*, *A* & not-*B*, not-*A* & *B*, and not-*A* & not-*B*. Estimates of the triples above showed that the participants making them didn't know how to compute probabilities. Their estimates yielded robust violations of the principle that the percentage probabilities in the JPD should sum to 100%. Likewise, individuals do not infer numerical probabilities unless the premises or the task prompts them to do so (Goodwin, 2014), which is just as well, given their incompetence with numerical probabilities.

3.2 The Model Theory of Probabilities

The model theory provides a mechanism for computing probabilities. Given this problem:

- There is a box in which there is a yellow card, or a brown card, or both.
- What is the probability that in the box there is a yellow card and a brown card?

individuals tend to respond with estimates of around 33%. Such "extensional" inferences follow from the proportion of models of possibilities in which the outcome occurs, and the disjunction yields three possibilities, of which only one satisfies the case in question (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). Similar models with numerical tags allow individuals to infer answers to problems of the following sort (Girotto & Gonzalez, 2002), in which the arithmetic is simple and calls for separate estimates of the denominator and the numerator of a probability:

> The chances that Pat has the disease are 4 out of 10. If she has the disease, then the chances are 3 out of 4 that she has the symptom. If she does not have the disease, then the chances are 2 out of 6 that she has the symptom. Pat has the symptom. Pat has __ chances of having the symptom, and among them she has __ chances of having the disease.

As the example shows, frequencies from observations are not necessary for correct inferences. The participants can envisage the solution in a series of updates to models of Pat's chances. The initial model establishes that Pat has 4 chances of having the disease and 6 chances of not having the disease. Likewise, her chances of having the disease divide into 3 chances in which she has the symptom and 1 chance in which she does not. And her chances of not having the disease divide into 2 chances in which she has the symptom and 4 chances in which she does not. In sum, the models of Pat's chances are as follows:

disease	symptom:	3
disease	¬ symptom:	1
¬ disease	symptom:	2
¬ disease	¬ symptom:	4

So, Pat has 5 chances (out of 10) of having the symptom, and given the symptom, she has 3 chances out of 5 of having the disease. So, with simple arithmetic, naïve individuals can use this "subset principle" (rather than Bayes' theorem) to estimate a conditional probability.

Many events in daily life have no frequencies of occurrence (e.g., the event that Trump will be reelected). Yet people in such cases can infer numerical probabilities. The mystery is where the numbers come from. The model theory offers a solution: they come from the proportions of models, not of the event itself, but of its occurrence in models of relevant evidence (Khemlani et al., 2015). So, for Trump's reelection, relevant evidence is, for example, that most U.S. presidents running for a second term haven't been reelected. This yields a model in which the smaller proportion of possibilities supports reelection. The model theory postulates that the intuitive system represents this magnitude in a nonnumerical way. It relies on an icon akin to one that nonnumerate individuals use to represent magnitudes (see, e.g., Carey, 2009). The greater the length of the icon, the greater the magnitude that it represents, so a probability is represented as a pointer on a "scale" extending from an origin representing impossibility to an end point representing certainty:

Subsequent evidence can shift the pointer one way or another, and its end point yields intuitive judgments, such as, "it's unlikely." The deliberative system can transform the icon into a numerical estimate: the probability is 40%. These two systems implemented in a computer program yield corroborated predictions about how individuals estimate the probabilities of assertions, including their propensity to make estimates that violate the probability calculus (Khemlani et al., 2015). The origins of these estimates, which go back to infancy, are in sampling from models of the world (e.g., Téglás et al., 2011).

Inferences of numerical probabilities should not flout the probability calculus. If they do, those who make them are vulnerable in principle to a "Dutch book"—a series of bets in which they are bound to lose money. Most of us, alas, often make irrational estimates of probabilities. Only rare experts are rational. A question that tests expertise in probabilities is about two events, *A* and *B*: what three sorts of estimate akin to the triples in section 3.1 fix the JPD in a way that always sums to 100%

whatever their numerical values? A correct answer is given below, but meanwhile readers should remember that the question stumps almost everyone.

Granted that most of us are not experts, our rationality can be assessed from our performance with simple extensional problems. But even in these tasks, we are vulnerable to illusory inferences:

There is a box in which there is at least a red marble, or else there is a green marble and there is a blue marble in the box, but not all three marbles.

What's the probability that there is a red marble and a blue marble in the box?

Most people say their joint occurrence is impossible, and so the probability is zero. Their estimate is predicted from the two mental models of the premise:

red marble
green marble blue marble

But these models fail to take into account the cases in which it is true that there is a red marble but false that there is both a green and blue marble. The latter conjunction is false, for instance, in case there is only a blue marble. So, there can be both a red marble and a blue marble in the box, and the zero estimate is an illusion.

In sum, probabilities are tricky. Their calculus is not intrinsic to human mentality. That is why it was necessary to invent it. As a philosopher once remarked, "Anyone who played dice in Roman times armed with the probability calculus would soon have won the whole of Gaul" (Hacking, 1975). Ignorance can underlie irrationality.

4. Induction and the Testing of Hypotheses

4.1 Constraints on Induction

Much of our reasoning is outside the scope of logic: it is inductive rather than deductive. And, as Hume (1748/1988, section IV) argued, induction relies on habit, not necessity. Theorists sometimes propose formal rules for induction, but they inherit the problem of logical form, and knowledge is crucial. If the first 10 people that you meet on a desert island are overweight men, knowledge guides you to induce that its inhabitants are more likely to be all obese than all men.

Knowledge can modulate the interpretation of connectives (e.g., Quelhas & Johnson-Laird, 2017), and so the boundary between deduction and induction is often unclear. The following inference is valid:

If it rained then the plants didn't die.

In fact, the plants did die.

So, it didn't rain.

Here's another example of what looks like the same sort of premises:

If it rained then it didn't pour.

In fact, it did pour.

Yet, you would be irrational to infer that it didn't rain. You know that in a meteorological context, *it pours* means that it rained hard. This knowledge modulates the interpretation of the conditional so that it refers only to a conjunction of two default possibilities: it rained but didn't pour, and it didn't rain. So the second premise contradicts the first premise. The example illustrates again the difficulty of pinning down logical form.

Beliefs underlie many inductions, which are therefore bound to be dangerous. Theorists have developed systems for updating beliefs (see section 5 of this volume). The trouble is that a set of beliefs can be consistent but irrational, because they do not correspond to reality. As a commentator on accidents at sea wrote, "Despite the increasingly sophisticated equipment, captains still inexplicably turn at the last minute and ram each other . . . they built perfectly reasonable mental models of the world, which work almost all the time, but occasionally turn out to be almost an inversion of what really exists" (Perrow, 1984, p. 230).

After a month, a new sort of light bulb fails in your living room. You draw an analogy: the bulbs are like some new flawed computer chips. You formulate an explanation: the bulbs' filaments contain an impurity. These processes are inductive, and they play a part in science and technology. For example, epidemiology was founded in part on John Snow's discovery of how cholera was communicated from one person to another. His discovery began with inductions: cholera is a single specific disease, its spread follows the trade routes—an infected mariner arriving in a port spreads the disease to those in contact with him. More puzzling was that it could leap over considerable distances. Snow inferred that "particles" of the disease in sewage contaminated drinking water. He was right: the particles were bacteria. The Wright brothers' invention of the airplane rested on numerous inductions. Maxim, the inventor of the eponymous gun, said, "Give us a motor, and we will very soon give you a successful flying machine." Wilbur Wright countered with an induction from the fate of glider pilots: if the engine fails and the pilot has no control of the aircraft, then it will crash and the pilot could be killed. The brothers' first goal was therefore to discover how to control a glider rather than to build a motor (for Snow's and the Wrights' inductions and deductions, see Johnson-Laird, 2006).

One constraint on inductions is similarity: similar causes have similar effects (Hume, 1748/1988, p. 80; Tversky & Kahneman, 1983). Another constraint is that physical contact implies causation—it is one cause of cholera's transmission. Yet, these constraints are those on which magical thinking and superstitions rely, and all cultures are susceptible: similarity yields homeopathic magic (take an antibiotic for a sore throat), and contact yields contagious magic (don't touch people with AIDS). There are too many possible hypotheses, and so any constraints are better than none. The only security in an induction comes from tests of its conclusion.

4.2 Testing Hypotheses

A test of an induction needs to determine whether its consequence is true or false. The most influential experimental paradigm for such studies is Wason's (1968) selection task. The experimenter chooses four cards from a pack in which, as the participants know, all the cards have a letter on one side and a number on the other side. They are laid out in front of the participant: *E F 2 3*. The participant's task is to select only those cards that, if turned over, would show whether a hypothesis about the four cards is true or false:

If there is an "E" on one side of a card then there is a "2" on the other side.

With an abstract hypothesis of this sort, participants tend to select the *E* and *2* cards or the *E* card alone. The selection of the *2* card is pointless, because nothing on its other side can refute the hypothesis. But what stunned psychologists and launched hundreds of studies was the participants' failure to select the *3* card. If it had an *E* on its other side, the hypothesis above would be false. Wason, his colleagues, and many others took this oversight to be irrational. Not everyone agreed. One view was that the probabilities of events related to everyday conditionals made it rational not to select the *3* card (see chapter 4.5 by Chater & Oaksford, this handbook). In contrast, the model theory's algorithm (Johnson-Laird & Wason, 1970a) predicts that the test of a general hypothesis needs to establish the possibility of examples and the impossibility of counterexamples (see section 2.1). So, selections of cards should be dependent on one another (e.g., the *2* card should tend to be selected with the *E* card, as an example of the hypothesis). It also predicts that the task and its contents should affect the likelihood that participants select potential counterexamples. A meta-analysis of over 200 experiments corroborated these predictions (Ragni, Kola, & Johnson-Laird, 2018).

In a repeated version of the selection task, participants had to make an efficient test of the generalization "All the triangles are white" about the shapes in two boxes (Johnson-Laird & Wason, 1970b). As the participants knew, one box contained 15 white shapes and the other box contained 15 black shapes. On each trial, they selected a shape from one of the two boxes to see whether or not it was a triangle. The participants started by selecting white shapes—potential examples of the hypothesis—but sooner or later switched to black shapes—potential counterexamples—and examined them all. The original selection task fooled them: they had just one opportunity to make a rational selection. In the repeated task, sooner or later they had the insight to select all potential counterexamples.

5. Abduction, Explanations, and Nonmonotonicity

5.1 Abductions That Resolve Inconsistencies

The author and a friend were sitting BC (before cell phones and corona virus) outside a café in Provence. We were waiting for two other friends, and we knew two things:

They had gone to pick up the car.

And if so then they would be back within 10 minutes.

So, we thought, they'll be back in 10 minutes. When they hadn't returned after 20 minutes, we realized that our conclusion was false. In logic, such a contradiction implies any conclusion whatsoever. Logic is "monotonic": a fact that contradicts an earlier valid inference does not call for its retraction. In daily life, however, it is rational to withdraw conclusions that facts refute, and so reasoning is "nonmonotonic" or "defeasible" (see chapter 5.4 by Gazzo Castañeda & Knauff, this handbook). William James (1907) wrote, "The new fact preserves the older stock of truths with a minimum of modification, stretching them just enough to make them admit the novelty" (p. 59). And this "minimalist" view has many defenders (see chapter 5.2 by Rott, this handbook). Yet, outside the café, we were concerned to figure out what had happened to our friends. We thought of various explanations: they had got lost (improbable), the police had stopped them (improbable), they had been in an accident (improbable), and they couldn't start the car (possible, because it had happened before). So, we reasoned, we had better sit tight and wait, because if we went in search of them and they returned to the café, they wouldn't know where we were. Sure enough, soon afterward, the car came spluttering into view, and we hopped in while its engine was running. It had needed a tow to start.

This anecdote exemplifies the model theory. Any inference in daily life is made in default of knowledge to the contrary. That is, conclusions can turn out to be false. In such cases, the primary task is to create an explanation that resolves the inconsistency between the fact and the premises. It can then guide the decision about what to do.

The creation of an explanation is known as "abduction." Its rational goal, as ever, is accuracy. The process is a sort of induction, because it goes beyond the premises and depends on knowledge. But it seeks more than an inductive generalization: it introduces new concepts in order to create an explanation. The process is triggered when a fact refutes an earlier inference. For example, the premises

If a cobra bit her then she'll die

A cobra bit her

elicit a mental model of a snake biting her and her death. But the fact

She did not die

is inconsistent with this model. And the conflict triggers an attempt to resolve the inconsistency. The whole nonmonotonic process is implemented in the mSentential program. Its first step is to compute the facts: a cobra bit her but she did not die. It searches its knowledge for such cases and their causes. The program's knowledge base contains several relevant possibilities in which a deadly snake bites someone who subsequently does not die:

- the person takes an antidote,
- a tourniquet blocks the poison, and
- someone sucks out the poison.

Human reasoners tend to assess the probability of each explanation. The program chooses one at random and then uses the same procedure again to search for *its* cause. If successful, the result is a causal chain, for example,

A cobra bit her and she used a tourniquet and the tourniquet blocked the poison and she did not die.

It also uses the first cause in the chain to construct a counterfactual description of what would otherwise have occurred:

If she had not used a tourniquet then she would have died.

Of course, the program is itself only a partial model of human reasoning. When you create explanations, you work with real models of the world (not strings of words), and you assess the relative probability of different putative explanations.

The theory predicts that individuals should prefer causal explanations to minimal revisions. Indeed, they rate a causal chain as more probable than either the cause alone or the effect alone (Johnson-Laird, Girotto, & Legrenzi, 2004)—a violation of minimalism and a fallacy in which a conjunction is judged to be more probable than its constituents (see Tversky & Kahneman, 1983). When individuals are asked what follows from inconsistent premises, they tend to explain the inconsistency. And they judge such explanations as more probable than minimal revisions to the premises that restore consistency (Khemlani & Johnson-Laird, 2011). When they formulate an explanation first, a striking phenomenon occurs. They find it harder to detect the inconsistency—they seem to have explained it away (Khemlani & Johnson-Laird, 2012).

A consistent set of beliefs may not correspond to reality, but an inconsistent set of beliefs *cannot* correspond to reality. The ability to detect inconsistencies is therefore a hallmark of rationality. Reasoners can assess consistency by trying to find a model of all the information they have. Such a model shows that the information is consistent; otherwise, it is inconsistent. Once reasoners have detected an inconsistency, they can use their knowledge to abduce causal models to explain the origins of the inconsistency. It is rational to try to do so, but no guarantee exists that the explanation is correct. And, on rare occasions, people fail to find any explanation whatsoever for an inconsistency. Yet, they outperform any existing computer program in creating explanations. The crux is that they know more and are better at retrieving pertinent information.

5.2 Abductions of Algorithms

Individuals who know nothing of programming can create simple informal algorithms. Deductions alone won't work, and probabilities are no use, either. Algorithms are an explanation of how to do something, and so reasoners need to introduce new concepts and to organize them in a temporal sequence. In a word, they *abduce* algorithms. For example, they can do so to rearrange the order of cars in a train on a railway track (Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013). The track runs from left to right on a computer screen, and it has one siding, which is entered from, and exited to, the left track. As an illustration, you might like to think how you would use the siding to rearrange the cars ABCD on the left side of the track so that they arrive in reverse order on the right side of the track. The more

moves (and the more cars in each move) that are called for to solve such a rearrangement, the harder the task is. A more substantial task, however, is to abduce an algorithm that solves such a rearrangement for a train with an arbitrary number of cars.

A computer program, mAbducer, creates such algorithms, which it can express in informal English. Here, for instance, is the algorithm it abduced to reverse the order of cars:

Move one less than the [number of] cars to the siding.
Move one car to the right track.
While there are more than zero cars on the siding,
    move one car to the left track,
    move one car to the right track.

The effect of the first move on ABCD on the left track is shown in the following diagram in which the square brackets demarcate the cars on the siding, and the dash shows that no cars are yet on the right side of the track.

    Move one less than the cars to the siding: A[BCD] –

The second instruction is:

    Move one car to the right track: – [BCD]A

Next, while there are more than zero cars on the siding, there is a loop of two instructions:

    move one car to the left track: B[CD]A
    move one car to the right track: – [CD]BA

The loop is repeated twice until the siding is empty, and the result is the cars in reverse order on the right track:

– [ ]DCBA

As mAbducer illustrates, the model theory postulates that people who are not programmers abduce informal algorithms using a kinematic model that simulates a solution to an instance of the problem. They observe what happens in the simulation and transform their observations into an informal description of the required algorithm. In principle, individuals could determine the number of iterations of a loop of moves by solving two simultaneous linear equations. But, in fact, they more often observe the conditions in which the loop halts in their simulations (e.g., when the siding is empty in the algorithm above). The program implements both sorts of loop (in both a programming language and English).

Experiments showed that most individuals follow the observational procedure—they use a so-called while-loop of the sort illustrated above (Khemlani et al., 2013). Even 10-year-old children can abduce informal algorithms for real toy trains of a small number of cars. In one experiment, they were not allowed to move the actual cars, and so they tended to make gestures corresponding to the moves their algorithms required. When they were prevented from gesturing, they were no longer so accurate in abducing algorithms (Bucciarelli, Mackiewicz, Khemlani, & Johnson-Laird, 2016). Gestures are an outward sign of inward simulation, and they reduce the load on working memory.

Different rearrangements vary in the difficulty of their abduction. For children and adults alike, it depends not on the number of cars that have to be moved but on the complexity of the algorithm. Various measures exist for this complexity, but a good proxy is the number of instructions in the algorithm. A rational algorithm should carry out the correct moves, and it should do so without unnecessary ones. It is tempting to say that it should be minimal (i.e., it should be of the shortest possible length in the given programming language). There is a problem, however: for all but the simplest algorithms, no way exists to prove that an algorithm is minimal (Chaitin, 1998).

6. Conclusions

Rationality can seem remote, abstract, and complex. The message of this chapter, however, is that it matters. What counts as rational is a recursive question, because its correct answer depends on rational methods. It goes beyond reasoning to apply to beliefs, actions, and social interactions. But, for reasoning, the view defended here is:

Rational reasoning constructs accurate models of the world based on knowledge or beliefs; infers their consequences for what is possible, probable, or certain; and uses these models to formulate explanations.

In short, right reasoning aims for truth, parsimony, and explanatory power. But, as the chapter showed, humans are fallible. They can fail to think about what's false in making deductions and succumb to illusory deductions. They can make an incorrect inference of the likelihood of a compound assertion, because the probability calculus is not part of their mental equipment. One test of expertise is my earlier question: which three sorts of estimate concerning two events, *A* and *B*, fix the values of their joint probability distribution and, regardless of their numerical values, always yield a JPD that sums to 100%? One correct answer is: the probability of *A*, the conditional probability of *B* given *A*, and the conditional probability of *B* given *not-A*. My survey of psychologists showed that few of them could answer this question correctly.

Although humans are adept at induction, no constraints can guarantee the truth of its outcomes. The heuristics that humans use are often irrational. They yield truths and trumpery, science and superstitions. One rational criterion, however, is that hypotheses should be open to refutation—for Popper (1959), this criterion separates science from nonscience. But naïve individuals in the selection task often focus on corroboratory examples of a hypothesis. In its repeated version, however, they do realize the need to ensure that counterexamples do not exist.

When a fact refutes the conclusion of a valid inference, logic tells us nothing—it licenses any conclusion whatsoever from a contradiction. In contrast, individuals innocent of philosophical niceties follow neither logic nor the advice to make a minimal revision of their beliefs. Instead, they search for an explanation that resolves the inconsistency. These explanations are not always minimal. The truth, as Oscar Wilde remarked, is seldom simple (and never pure). One domain, however, does call for minimalism: the abduction of algorithms. Like so many ideals, alas, no way exists guaranteed to reach the shortest possible algorithm.

A common error in all sorts of reasoning is to overlook a possibility. Such failures are not inevitable, and the model theory suggests something that can be done to reduce them. You can learn to be more rational. You might imagine that you have to learn logic. It takes time—at least a semester's course for the rudiments—and it fails to generalize to novel sorts of inference (Cheng, Holyoak, Nisbett, & Oliver, 1986). In contrast, a good method for improving reasoning should be quick to learn, efficacious, and practical. One method meets these goals. Its single instruction is:

Try to construct all the possibilities consistent with the given information.

With pencil and paper, you can list them in separate columns, adding a column for each new possibility and crossing out a column if a premise refutes its possibility. It takes only a few minutes to learn this "model method," which Victoria Bell devised (see Johnson-Laird, 2006, p. 288). It speeds up reasoning and increases its accuracy (from 66% correct to 95% in one study). Once people have acquired the method, they can imagine the columns of possibilities and no longer need pencil and paper. It still works. And the method may be adaptable to other sorts of reasoning beyond deduction.

To answer the two questions in my opening paragraph: right reasoning should yield accurate models of the world, and human reasoning is rational in principle but often irrational in practice. Yet, humans realize their own shortcomings, and they are rational enough to develop tools that can help them. Contrary to Wilde's epigraph at the head of this chapter, psychology grapples with the irrational and even dispels it . . . sometimes.

References

Adams, E. W. (1998). *A primer of probability logic*. Stanford, CA: Center for the Study of Language and Information.

Boole, G. (1854). *An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities*. London, England: Macmillan.

Bucciarelli, M., Mackiewicz, R., Khemlani, S. S., & Johnson-Laird, P. N. (2016). Children's creation of algorithms: Simulations and gestures. *Journal of Cognitive Psychology*, *28*, 297–318.

Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.

Chaitin, G. J. (1998). *The limits of mathematics*. Singapore: Springer.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, *18*, 293–328.

Craik, K. J. W. (1943). *The nature of explanation*. Cambridge, England: Cambridge University Press.

Einstein, A. (2004). *Relativity*. New York, NY: Barnes & Noble. (Original work published 1920)

Girle, R. (2009). *Modal logics and philosophy* (2nd ed.). London, England: Routledge.

Girotto, V., & Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning. *Cognition*, *84*, 353–359.

Goodwin, G. P. (2014). Is the basic conditional probabilistic? *Journal of Experimental Psychology: General*, *143*, 1214–1241.

Hacking, I. (1975). *The emergence of probability*. Cambridge, England: Cambridge University Press.

Hempel, C. G. (1945). Studies in the logic of confirmation, Parts I and II. *Mind*, *54*, 1–26, 97–121.

Hinterecker, T., Knauff, M., & Johnson-Laird, P. N. (2016). Modality, probability, and mental models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1606–1620.

Hume, D. (1988). *An enquiry concerning human understanding* (A. Flew, Ed.). La Salle, IL: Open Court. (Original work published 1748)

James, W. (1907). *Pragmatism—a new name for some old ways of thinking*. New York, NY: Longmans.

Jeffrey, R. C. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York, NY: McGraw-Hill.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (2006). *How we reason*. New York, NY: Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*, 640–661.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*, 201–214.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition*, *193*. doi:10.1016/j.cognition.2019.04.019

Johnson-Laird, P. N., & Wason, P. C. (1970a). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*, 134–148.

Johnson-Laird, P. N., & Wason, P. C. (1970b). Insight into a logical relation. *Quarterly Journal of Experimental Psychology*, *22*, 49–61.

Keene, G. B. (1992). *The foundations of rational argument*. Lewiston, NY: Edwin Mellen Press.

Khemlani, S. S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*, 276–288.

Khemlani, S. S., & Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica*, *139*, 486–491.

Khemlani, S. S., & Johnson-Laird, P. N. (2017). Illusions in reasoning. *Minds and Machines*, *27*, 11–35.

Khemlani, S. S., Lotstein, M., & Johnson-Laird, P. N. (2015). Naive probability: Model-based estimates of unique events. *Cognitive Science*, *39*, 1216–1258.

Khemlani, S. S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences*, *110*, 16766–16771.

Laplace, P. S. de. (1951). *Philosophical essay on probabilities*. New York, NY: Dover. (Original work published 1820)

Lewis, D. (1981). Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, *10*, 217–234.

Nicod, J. (1950). *Foundations of geometry and induction*. New York, NY: The Humanities Press.

Northcutt, W. (2002). *The Darwin awards: Evolution in action*. New York, NY: Plume.

Oaksford, M., Over, D., & Cruz, N. (2019). Paradigms, possibilities, and probabilities: Comment on Hinterecker, Knauff, and Johnson-Laird (2016). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 288–297.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. New York, NY: Basic Books.

Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Quelhas, A. C., & Johnson-Laird, P. N. (2017). The modulation of disjunctive assertions. *Quarterly Journal of Experimental Psychology*, *70*, 703–717.

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses. *Psychological Bulletin*, *144*, 779–796.

Sperber, D. (1995). Intuitive and reflective beliefs. *Mind & Language*, *12*, 67–83.

Steingold, E., & Johnson-Laird, P. N. (2002). Naive strategic thinking. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 845–849). Fairfax, VA: Erlbaum.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*, 1054–1059.

Tversky, A., & Kahneman, D. (1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 292–315.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Yardley, I. (2014). *Ninety seconds at Zeebrugge: The Herald of Free Enterprise story*. Stroud, England: History Press.

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