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

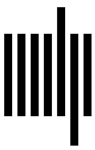
The Handbook of Rationality

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

10.6 Adaptationism: A Meta-Normative Theory of Rationality

Leda Cosmides and John Tooby

Summary

Human cognition is often compared—unfavorably—to normative theories of rationality from mathematics, logic, economics, or philosophy. But what justifies *these* normative theories as the proper standard for assessing the rationality of an evolved computational system? The adaptationist program in evolutionary biology provides a meta-normative theory of rationality that is appropriate for evolved organisms. The functional design of our cognitive architecture was built by natural selection, to solve information-processing problems that are strange, exact, and nonintuitive. Theories of adaptive function—task analyses of these problems—provide normative standards of good design for assessing the rationality of human cognition. With five case studies, we show how this approach can reveal sophisticated cognitive mechanisms that would otherwise remain undetected. At the same time, these cases illustrate pitfalls of studying reasoning and choice without reference to the ancestral problems and environments that selected for their design.

1. The Paradox of Human Reasoning

How the mind works can be illuminated by comparing human cognition to standards of good design specified by a normative theory. But which standards are appropriate for an evolved organism? What counts as a rational inference or choice for animals like us, whose minds were designed by natural selection?

The first normative theories used in cognitive psychology were those deemed rational by mathematicians, scientists, economists, and philosophers. Deductive and inductive reasoning were compared to standards of rationality drawn from first-order logic, Bayes' theorem, or statistical principles. The manner in which people evaluate hypotheses was compared to Popper's standard: look for violations, not confirmation. Decision making was compared to Neyman–Pearson decision theory, expected utility theory, or, when choosing among bundles of goods, the axiom of transitivity (e.g., GARP).

In this research tradition, reasoning errors are defined as having occurred whenever human judgment departs from what these normative theories dictate. In their seminal studies of judgment under uncertainty, Kahneman and Tversky (1982) articulated this standard: "The presence of an error in judgment is demonstrated by comparing people's responses either with an established fact (e.g., that the two lines are equal in length) or with an accepted rule of arithmetic, logic, or statistics" (p. 123).

When cognitive performance was assessed by these normative standards, the rationality of human reasoning, judgment, and decision making was found wanting. Psychologists found systematic errors in deductive reasoning and hypothesis testing, violations of probability theory, preference reversals, and "excess" altruism. These "errors" were attributed to a flotilla of heuristics, biases, and emotions—cognitive processes thought to be fast, automatic, unconscious, emotional, associative, and/or stereotypic. This heterogeneous collection of processes was called "System 1." Rational thinking was attributed to "System 2": slow, effortful, conscious deliberations that carry out calculations and logical inferences (Kahneman, 2011).

This research program has produced a formidable paradox. Reviews of human reasoning research are customarily presented as lengthy catalogs of errors, fallacies, biases, heuristics, and emotions. Textbooks organize their discussions around the seemingly endless ways in which human reasoning and decision making depart from the normative ideals of rationality used in science, mathematics, economics, and philosophy. Yet evolved reasoning systems—human and nonhuman minds alike—negotiate the complex natural tasks of their world with a level of operational success far surpassing that of the most sophisticated existing artificial intelligence (AI) systems. It is trivial to equip AI systems with algorithms that implement statistical decision theories, inferences of first-order logic, and other formal methods of rational

inference. They can search vast problem spaces in seconds, store massive databases for analysis, and perform lightning calculations. But organisms perform better than these computational systems on virtually every natural inferential problem that has been carefully investigated—the induction of grammar and word meanings, speech perception, vision, color constancy, recognizing objects and making inferences about their interactions, and inferring the beliefs, desires, and behavior of other people, to name a few.

What is the resolution to this paradox? From the perspective of evolutionary biology, the problem is not that our thinking is irrational. The problem is how psychologists have been defining rationality and testing for its presence.

2. A Meta-Normative Theory of Rationality

Let us take one step back and ask a more fundamental question. Why identify rationality with adherence to normative theories from logic, mathematics, economics, and philosophy? What normative theory justifies the choice of these particular normative theories?

In common parlance, a choice, process, or behavior is considered *rational* when it is well designed for achieving a goal, and *irrational* when poorly designed for that function. It would be irrational to eat charcoal instead of fruit to sate hunger (but rational if you had just ingested a poison); it would be irrational to travel to a new location by randomly zigzagging instead of following a straight-line path (but rational if you are evading a predator); it would be irrational to converse out loud while alone (but rational if you are rehearsing a role). There is no goal-independent definition of what counts as rational.

The specialized (and historically recent) tools of scientists, mathematicians, economists, and philosophers were developed for very specific goals: producing knowledge, understanding choice, and guiding behavior in mass societies. From the 1400s to the present, increasing literacy, more reliable data storage technologies, and the decreasing cost of communication and travel have changed the world beyond anything known by our hunter–gatherer ancestors. Because of these developments, hypotheses and data became more widely shared, debated, vetted, and stored, enabling the cultural accumulation of knowledge in ever-larger populations; markets arose in which millions of anonymous individuals cooperated in complex networks; people from locations with different moral norms were brought into contact, in ever-widening circles of interaction. Applying standards of rationality created by scientists, mathematicians, economists, and philosophers is useful for solving the evolutionarily novel problems created by these conditions, such as producing scientific knowledge, understanding market dynamics, and, perhaps, fostering cooperation in the modern world of mass societies and global contact. But what justifies privileging methods developed for these modern goals as *the* metric against which human rationality is measured?

Nothing. From an evolutionary perspective, these are the wrong goals against which to measure human rationality, because they played no causal role in selecting for the design of the mind. To decide whether a mechanism of human reasoning or decision making is rational—whether it is well engineered for producing a goal or outcome—we need to know what adaptive problem that mechanism was designed (by natural selection) to solve, and what kind of information was available for solving it in the environments that selected for its design.

3. The Adaptationist Program in Evolutionary Biology Provides a Sound Meta-Normative Theory of Rationality

From the molecular machinery of a single cell to the information-processing architecture of the human visual system, the most sophisticated engineering on Earth is found in organisms and was built by natural selection.

Organisms are composed of many machines: systems organized to solve a problem. *These systems evolved to solve problems in ways that improved the ability of the organism's ancestors to produce offspring.* Every piece of organic machinery—the retina, the heart, the lungs—acquired its intricate functional design over deep time, as a downstream consequence of the fact that organisms reproduce themselves.

3.1 Natural Selection Retains Designs That Promoted Their Own Reproduction

When individuals reproduce, replicas of their organic machinery develop in their offspring. But replication is not error-free: chance mutations introduce changes into the design of organic machines, altering their features. These entropic changes usually disrupt the efficiency with which this machinery solves problems, thus interfering with the mutant offspring's ability to produce offspring of its own. Because individuals with the newly modified, but now defective, design produce fewer offspring, on average, than those with the (more efficient) standard design, the defective design eventually disappears from the population—a case of negative feedback.

Occasionally, a mutation will improve a machine's operation in a way that promotes the reproduction of individuals with that mutation—and, therefore, the reproduction of that design. Such improved designs (by definition) cause their own increasing frequency in the population—a case of positive feedback. This increase continues until (usually) the modified design out-reproduces, and thereby replaces, all alternative designs in the population, leading to a new species-standard design. The population- or species-standard design has taken a step "uphill" toward a greater degree of functional organization for reproduction than it had previously.

Over the long run, down chains of descent, this feedback cycle—natural selection—pushes designs through state-space toward increasingly well-engineered functional arrangements. These arrangements are *functional* in a specific sense: the elements are well organized to cause their own reproduction in the environments in which the species evolved.

3.2 Adaptive Problems and Evolved Solutions

Enduring conditions in the world that create reproductive opportunities or obstacles constitute *adaptive problems*. Examples include the presence of predators, the possibility of sharing food to pool foraging risk, or the existence of a cognitive mechanism that could be repurposed for more efficient foraging. Enduring relationships of this kind constitute reproductive opportunities or obstacles in the following sense: *if* the organism had a new property that interacted with these conditions in just the right way, *then* this property would cause individuals who have it to produce more offspring that live to reproductive maturity, relative to those with alternative designs. These reproductive opportunities and obstacles can be thought of as *problems*. A property is a *solution* to such a problem when it allows organisms with this property to take advantage of prevailing conditions, where "advantage" means a reproductive advantage.

Adaptive problems have two defining characteristics. First, they are conditions or cause-and-effect relationships that many or most individual ancestors encountered, reappearing again and again during the evolutionary history of the species. Second, they are that subset of enduring relationships that could, in principle, be exploited by some new property of an organism to increase its reproduction or the reproduction of its relatives (who have a high probability of having inherited the same mutation).

An enduring adaptive problem constantly selects for design features that promote the solution to the problem. Over evolutionary time, more and more design features accumulate that fit together to form an integrated structure or device that is well engineered to solve that adaptive problem. Such a structure or device is called an *adaptation*. An adaptation may have many beneficial effects, but solving the adaptive problem that selected for its design is its *function*. The function of the heart is to pump blood; the function of the liver is to detoxify poisons; the function of the retina is to detect photons and transduce them into neural signals for vision. Adaptive problems that required information processing for their solutions selected for neural circuitry organized to compute these solutions: computational adaptations.

Adaptationism is the name of the research program that explores how natural selection functionally organizes the designs of organisms. Adaptationists start with a careful analysis of an adaptive problem—including information that would have been available in ancestral environments for solving that problem. *A careful task analysis of an adaptive problem serves as a normative theory of good design:* it identifies the problem, allowing one to see what counts as a good solution to it. Because a computational adaptation should be composed of algorithms and representations that are well engineered for solving the problem that selected for its design, a good theory of adaptive function suggests hypotheses about the design of the mind. These can be tested empirically, revealing computational operations that were previously unknown.

The adaptationist perspective provides principled, nonarbitrary criteria for judging rationality. If a choice, process, or behavior is *rational* when it is well designed for achieving a goal, then a computational adaptation that generates inferences or choices is rational when its features are well tailored for solving an adaptive problem faced by the hunter–gatherers from whom we are descended.

The study of reasoning and choice looks very different when viewed from this perspective. Many fallacies disappear, and well-designed reasoning appears, when cognitive performance is compared to theories of adaptive function: normative standards of good design derived from evolutionary biology, behavioral ecology (of hunter–gatherers, nonhuman primates, and other animals), and studies of ancestral environments (from paleoanthropology, archaeology, physics, geology, botany, zoology, and other sources). Using five case studies of reasoning and choice, we illustrate pitfalls that arise from the failure to apply adaptationist thinking. Cases 3–5 tap theories of adaptive function that are normative in evolutionary biology but have no counterpart in traditional normative theories.

4. Case 1: To Discover a Well-Designed Reasoning Mechanism, You Need to Present Stimuli in an Ecologically Valid Format

Behavioral ecologists have sophisticated models of the adaptive problems animals face when foraging. Because efficient foraging requires animals to make judgments under uncertainty, these models incorporate elements of probability theory. Behavioral ecologists typically find that insects, birds, and other animals behave like good Bayesians when they make foraging decisions (e.g., Real, 1991; Stephens & Krebs, 1986). Yet psychologists thought that the human mind was "too limited" in capacity to do the same (Kahneman, Slovic, & Tversky, 1982). Why the difference?

Computational adaptations should be *ecologically rational*: designed to work well in the ecological circumstances that selected for their design. Consider, for example, how well your visual system maintains color constancy given normal variations in light cast by the sun: your green car looks green all day, even at sunset, when it is bathed in "red" (long-wavelength) light. Yet these sophisticated mechanisms fail—your green car looks brown—in the ecologically novel spectrum cast by sodium vapor streetlights. When psychologists first started to study Bayesian reasoning, the stimuli they used were the cognitive equivalent of sodium vapor lights: probabilities of single events (e.g., there is a 4% chance you will find an apple tree in this orchard) and normalized frequencies (36% of the trees are cherry trees). These data formats are the recent cultural product of modern data gathering and statistical techniques.

When asked to compute a conditional probability based on such data (e.g., What is the chance that a tree with red fruit is an apple tree?), people fail spectacularly. But they succeed when given natural frequencies: the *absolute* frequencies of events *as you encounter them* in the world (Gigerenzer & Hoffrage, 1995, 1999). Our minds automatically encode natural frequencies, which were the only kind of probability data available in the ecology of our hominin ancestors (Cosmides & Tooby, 1996). Imagine an orchard with 125 fruit trees—apple, cherry, lemon, and pear. On a stroll, you encounter 50 trees with red fruit: 5 apple trees and 45 cherry trees. Given these natural frequencies, most people realize that 5 out of 50 trees with red fruit are apple trees—the conditional probability that a tree with red fruit is an apple tree (hits / (hits + false alarms)). Note that the low base rate of apple trees (4%) is implicit in the small number of them that you encounter as you walk through the orchard (thereby taking a "natural sample"; Kleiter, 1994). Adaptations for making probability judgments can safely ignore normalized data about base rates (as people famously do) if they are designed to use natural frequencies derived from natural samples. (On the fit between cognitive mechanisms and their task environment, see chapter 8.5 by Hertwig & Kozyreva, this handbook.)

Our mental mechanisms do produce judgments that match normative standards from probability theory—when they are fed data in the format they were designed to use. The practical relevance for modern environments is clear: a simple change in how data are communicated can help patients and policy makers make rational decisions about risk (Gigerenzer, 2014).

5. Case 2: Judgments That Look Like Normative Violations May Be Very Well Designed for Solving a Recurrent Adaptive Problem—One the Scientist Is Not Considering

In choosing between risky and sure options, many people behave as if "losses loom larger than gains": they are willing to take a bigger risk to avoid *losing* $100 than to *gain* the same amount. To appreciate the puzzle, imagine there is an outbreak of a disease that will kill 600 people if nothing is done. Which of two programs to combat the disease would you favor?

> If program A is adopted, *400 people will die*. If program B is adopted, there is 1/3 probability that *nobody will die*, and 2/3 probability that *600 people will die*. (italics added)

Expected utility theory predicts indifference, because both programs have the same expected value. With program A, 400 people die and 200 survive; program B produces the same result, on average. But Tversky and Kahneman (1981) found that people are not indifferent: when the options were framed as lives lost, most people ($\sim$ 80%) chose the risky option, program B.

Even more puzzling: this preference reversed—most people ($\sim$ 70%) chose the sure option, program A—when the same options were framed as gains:

> If program A is adopted, *200 people will be saved*. If program B is adopted, there is 1/3 probability that *600 people will be saved,* and 2/3 probability that *no people will be saved*.

Why? The loss and gain frames express logically and mathematically identical situations. Preference reversals are considered irrational: they violate the transitivity axiom of expected utility theory.

Because expected utility theory cannot explain these choices, psychologists and economists proposed that people have a stable taste: an aversion to loss. But a

normative theory from behavioral ecology explains these results—and correctly predicts when they will flip.

5.1 Risk-Sensitivity Theory, an Alternative Analysis of Rational Decision Making

Consider a bird who is deciding to forage on one of two patches. Both patches yield about 200 seeds per day of foraging—they have the same expected value—but they differ in variance. The low-variance patch yields close to 200 seeds each day. The yield of the high-variance patch varies wildly from one day to the next, from 50 to 400 seeds.

Now consider the bird's needs. If the bird needs to find 100 seeds today to live to tomorrow—a number below the expected value—the safest bet is to forage on the low-variance patch. But this is a bad bet if the bird needs 300 seeds to live another day: a rational bird will forage on the high-variance patch when its need is above the expected value. According to risk-sensitivity theory, a normatively correct decision takes three variables into account: (1) the expected value of each option, (2) the outcome variance associated with each option, and (3) the decision maker's need level. A rational decision system will be designed to minimize the probability of an outcome that fails to satisfy one's need (Stephens & Krebs, 1986).

To test risk-sensitivity theory in humans, the experimenter must vary the minimum need level before subjects make a decision. When no need level is specified, the subject is free to fill in the blank any way she wishes. For monetary gambles, maintaining the status quo is as good a minimum need level as any; a rational individual who does not want to fall below the status quo will avoid the risky option and choose the sure one. But this is not because she has an aversion to loss—a stable taste—that is independent of context.

Like birds, people choose rationally when their minimum need level is specified (Mishra & Fiddick, 2012; Rode, Cosmides, Hell, & Tooby, 1999). When it is *above* the expected value, people *prefer* the risky option. They choose the sure/low-risk option—thereby appearing to be loss averse—only when they need less than the expected value. Framing affects choice on the disease problem by changing the minimum number of lives people feel they *must* save: saying that 400 people will die leads people to set a higher threshold than saying that 200 will survive (Mishra & Fiddick, 2012). As a bonus, the same theory shows that "ambiguity avoidance" is also a myth (Camerer & Weber, 1992). Ambiguity is usually interpreted as "risky," but when the context implies it is the lower-variance option, people *prefer* the ambiguous option (Rode et al., 1999).

A tangle of results that look irrational when compared to traditional normative theories turn out to be rational responses to a problem that scientists were not considering.

6. Case 3: Cues of a Reasoning System's Proper Domain—the Context for Which It Evolved—May Be Necessary to Activate Its Procedures. To Know Its Proper Domain, One Needs to Correctly Characterize the System's Adaptive Function

An adaptation's *proper domain* is the information that the mechanism was designed by natural selection to process (Sperber, 1994). Ground squirrels, for example, evolved to produce alarm calls when they see a hawk overhead. The sight of a hawk is a cue that the squirrel is in danger from raptors *now*. This cue elicits the correct alarm call from the squirrel's repertoire (snakes and predatory cats elicit different calls). The approach of predatory birds is the adaptation's proper domain.

The raptor call is also activated when zoologists fly a drone with the silhouette of a hawk overhead. The alarm call's *actual* domain is larger than its proper domain: it consists of all cues that activate the call, real hawks (proper domain) and hawk silhouettes alike. In studying a reasoning mechanism in humans, experimenters need to consider (i) the mechanism's evolved function, to ascertain its proper domain, and (ii) whether the experimental stimuli present cues from its actual domain—ones sufficient to activate the mechanism. Mercier and Sperber (2011) have argued that most studies finding poor logical reasoning lack cues to the mechanism's proper domain: devising and evaluating arguments intended to persuade or dissuade another person.

Eliciting poor logical reasoning from people is easy: give them a problem, without context, that requires them to produce a valid inference using modus tollens (MT). MT is an inference rule from first-order logic. Given a conditional rule, such as "If the pipe was fixed, then the bathroom floor will be dry in the morning," and a premise, such as "The bathroom floor is wet this morning," it is logically correct to infer "The pipe was not fixed" (more generally: *If P then Q*; *not-Q*; therefore *not-P*). It is trivial to program a computer to produce an MT inference. Yet people fail to do so about 75% of the time (Rips, 1994; Wason & Johnson-Laird, 1972).

But do our minds lack an MT inference rule? Or do indicative conditional rules stripped of context fail to activate it? The mind does seem to implement certain rules of first-order logic (e.g., people usually apply the modus ponens rule, correctly inferring *Q* from *If P then Q* and *P*). If the human mind has a system designed for

reasoning logically, what adaptive problem did it evolve to solve? Mercier and Sperber (2011) have proposed that conscious, deliberative reasoning—including logical reasoning—evolved to solve adaptive problems that arise during communication.

6.1 Biologists Have a Normative Theory of Communication

A seminal paper on the evolution of communication by Krebs and Dawkins (1984) changed how biologists study animal communication. A communication system will not evolve unless it confers a net benefit on both senders and receivers—if a signal benefits only senders, receivers will not evolve mechanisms to decode it. But once a signaling system has evolved, situations will arise in which senders can manipulate receivers, to their own advantage, by sending deceptive signals. If acting on deceptive signals is costly to the fitness of the receiver, selection will favor adaptations in receivers to detect which signals are deceptive. As receivers get better at distinguishing honest from deceptive signals, selection will favor adaptations in senders for producing deceptive signals that are more difficult to detect, and so on. A coevolutionary arms race ensues over generations: receivers get better and better at detecting deceptive signals, and senders get better and better at making deceptive signals resemble honest ones. By contrast, the fitness interests of receivers and honest signalers converge; this favors the evolution of honest signals that are difficult for a deceptive sender to fake (Higham, 2014).

Language presents additional problems, however, because sending a deceptive message is no more costly (in words produced) than sending an honest one. Detecting the veracity of a signal requires logical reasoning, according to Mercier and Sperber, so that receivers can maintain "epistemic vigilance." While the receiver is evaluating the sender's claim, it is stored in a special data format—a meta-representation—which is decoupled from the receiver's semantic memory (her database of knowledge). Logical reasoning is necessary for her to detect whether the claim contradicts other claims made by the sender, claims made by other people, and facts she already knows. This evaluation process is particularly important, they claim, in arguments intended to persuade the receiver to act on beliefs she does not yet hold or dissuade her from actions she wants to take.

6.2 Argumentative Context: A Cue That Activates Logical Reasoning

When the sender and receiver already agree, there is no argument—no attempt to persuade or dissuade. If they disagree and an argument ensues, there are two possibilities: (i) one (or both) of them are misinformed (and they need more accurate information), or (ii) a deceptive message is being sent, intended to manipulate the receiver into doing something against her interests that benefits the sender. Both call for epistemic vigilance by the receiver: the claims must be evaluated for their truth value. The presence of an argument is, therefore, a cue that should activate logical reasoning in the receiver. It should also activate logical reasoning in the sender, who must devise an argument that will persuade the receiver.

This has empirical implications. Divorced from the context of an argument, the MT inference may not be activated. But it can be easily activated in an argumentative context. Let's say I am arguing with the plumber, who claims that he fixed the leaky pipe in my powder room. I doubt him, and deny his claim by saying, "Oh yeah? We'll see! *If the pipe was fixed, then the bathroom floor will be dry in the morning.*" When I discover that the bathroom floor is wet in the morning, we will both conclude that the pipe was not fixed: we easily make the MT inference.

Mercier and Sperber review evidence in favor of their hypothesis that an argumentative context is the proper domain for reasoning, logical and otherwise. Their claim is not that reasoning will be flawlessly logical in all arguments: an argumentative context will activate logic when the reasoner must be epistemically vigilant, but confirmation bias and other infelicities will emerge when sound reasoning would undermine one's attempt to persuade. In their words, "In all these instances traditionally described as failures or flaws, reasoning does exactly what can be expected of an argumentative device: Look for arguments that support a given conclusion, and, ceteris paribus, favor conclusions for which arguments can be found" (Mercier & Sperber, 2011, p. 57).

7. Case 4: A Cognitive Adaptation for Reasoning May Be Specialized for a Specific Domain and, Therefore, Equipped with Procedures That Are Content Rich Rather Than Content Free

Normative theories of rationality from logic, mathematics, and economics typically posit content-free reasoning procedures: ones that operate uniformly on information from every domain. Such procedures—*domain-general* ones—do exist in the human mind (automatic frequency computation is an example), but they cannot solve even routine adaptive problems by themselves (Cosmides & Tooby, 1987; Tooby & Cosmides, 1992). Domain-specialized reasoning systems were also required. These are equipped with content-rich concepts, inference procedures, and decision rules, ones that are superbly engineered for

producing fitness-promoting inferences in one ancestral domain but do not apply outside it (e.g., concepts like *belief* and *desire* are useful for predicting the behavior of people, but not rocks). Discovering these systems requires careful analysis of an adaptive problem to see what counts as a functional (i.e., normative) solution.

The evolution of cooperation for mutual benefit—social exchange—poses exacting problems, which have been modeled extensively. In humans, these problems are solved by a functionally specialized reasoning system that deploys content-rich representations and procedures. Its features were revealed by experiments that tested hypotheses derived from evolutionary game theory—hypotheses that were constructed in advance of collecting any data.

7.1 Evolutionary Game Theory Is a Source of Normative Theories

Game theory is a tool for analyzing strategic social behavior—how agents will behave when they are interacting with others who can anticipate and respond to their behavior (chapter 9.3 by Alexander, this handbook). Economists use it to analyze how people respond to incentives present in the immediate situation. Their models typically assume rational actors, who calculate the payoffs of alternative options (anticipating that other players will do likewise) and choose the option most likely to maximize their short-term profits (but see Hoffman, McCabe, & Smith, 1998).

Evolutionary biologists also adopted game theory as an analytic tool, but with a twist (Maynard Smith, 1982). Evolutionary game theory does not assume economically rational agents who can reason about the reasoning of other agents via "backward induction." It can be usefully applied to cooperation among bacteria or fighting in spiders. It is used to model interactions among agents endowed with well-defined decision rules that produce situationally contingent behavior. Although these decision rules are sometimes called "strategies" by evolutionary biologists, no conscious deliberation by bacteria (or humans) is implied (or ruled out) by this term. Sometimes results are derived analytically; in more complex cases, agent-based simulations of natural selection are used.

Whether the decision rules being analyzed are designed to regulate foraging, fighting, or cooperating, the immediate payoffs of these decisions, in food or resources, are translated into the currency of offspring produced by the decision-making agent, and these offspring inherit their parent's decision rule. In evolutionary game theory, a decision rule or strategy that garners higher payoffs leaves more copies of itself in the next generation than alternatives that garner lower payoffs. By analyzing the reproductive consequences of alternative decision rules over generations, evolutionary biologists can determine which strategies natural selection is likely to favor and which are likely to be selected out. This source of normative theories can be tapped for any domain in which organisms make consequential choices.

The evolution of cooperation has been extensively modeled using these tools, with clear results. Strategies that indiscriminately provide benefits to others—including those who do not reciprocate—are eventually eliminated from the population. Because they incur reproductive costs without compensating benefits, they are outcompeted by designs that cheat—that accept benefits without reciprocating them. Selection favors strategies that cooperate conditionally: ones that cooperate with other cooperators and withdraw cooperation from cheaters (e.g., Axelrod, 1984; Trivers, 1971). A *cheater* is an agent endowed with decision rules that accept benefits offered by other agents without satisfying their requirements. Innocent mistakes do not reveal a cheater; a cheater is an agent who violates a social exchange agreement by design (by virtue of its decision rules), not by accident.

This analysis carries many implications about how a system must be designed to implement a strategy for conditional cooperation. The most straightforward: for social exchange to evolve, agents must have mechanisms that make them very good at searching for information that would reveal cheaters. A content-free logic, deontic or otherwise, cannot do the job (e.g., Cosmides & Tooby, 2008).

7.2 Information Search: Looking for Cheaters versus Looking for Logical Violations

The philosopher Karl Popper proposed a normative standard for evaluating hypotheses: look for violations, not confirmation. "All swans are white" is violated by finding a single black swan, no matter how many white swans you have so far observed. (Realizing this requires the MT inference.) Peter Wason developed his four-card selection task to find out if people are natural falsificationists, who look for cases that could violate a hypothesis that is presented as a conditional rule (*If P then Q*). His research suggests that we are not (Wason & Johnson-Laird, 1972).

Imagine, for example, that you are given incomplete information about four birds—two of unknown color (a swan and a parrot) and two of unknown species (one black and one white). Which birds should you investigate further to see if any of them violate the rule "If a bird is a swan, then it is white"? Most people want to investigate the swan, to learn its color. That is logically correct: discovering a black swan would violate the rule. But that implies you should also investigate the black

bird, to learn its species: it too might be a black swan. Most people fail to investigate the black bird, and many want to investigate the white one (unnecessary: white birds—swan, dove, parrot—cannot violate the rule). Only cases of *P* & *not-Q* (here, swans that are not white) can violate *If P then Q*. Most people recognize this when asked, but they do not spontaneously use logic to *search* for violations of indicative rules. Fewer than 25% of people seek information about *P*, *not-Q*, and no other case.

By contrast, 65% to 80% of people successfully look for violations—cheaters—when the rule involves social exchange. The mind interprets a conditional rule as a *social contract* when it expresses an agreement to cooperate for mutual benefit, for example, "If you borrow my car, then you must fill the tank with gas" (or, more generally, "If you accept *benefit B* from agent *J*, then you must satisfy *J*'s *requirement R*"). It becomes obvious that one needs to investigate the guy who borrowed the car (*P*) and the one who did not fill the tank (*not-Q*).

7.3 A Content-Rich Adaptive Logic

Wason tasks involving social exchange activate a cognitive adaptation that evolved for detecting cheaters. It is part of a computational system that is specialized for reasoning about social exchange (for a review of the evidence, see Cosmides & Tooby, 2015). The social exchange system dissociates, both functionally and neurally, from reasoning systems that are activated by content tapping other domains (including precautionary rules, which are so similar to social contracts that most theories do not distinguish them). It represents social contracts using content-rich proprietary concepts (e.g., *agent*$_i$, *benefit to agent*$_i$, *requirement of agent*$_j$, *obligation*, *entitlement*, *cheater*). Its procedures operate on these representations, producing inferences appropriate to social exchange that are not licensed by content-free logics (deontic or otherwise). Its cheater detection mechanism attends to information that would reveal cheaters, whether the resulting answer is logically correct or not. And that mechanism looks for *cheaters*—innocent mistakes do not elicit violation detection. The design features of this content-rich system are normatively correct: they are precisely tailored for their adaptive function.

8. Case 5: Choices That Deviate from Economic Rationality May Be the Most Adaptive, Fitness-Promoting Strategy—Not a "Bias"

Economists have built models in which rational individuals and firms behave "as if" they were maximizing profits (or, more accurately, utility functions). A utility function is a mathematical device for understanding the dynamics of markets in which millions of anonymous individuals cooperate for mutual benefit: it transforms changes in inputs (e.g., the price of corn) into changes in outputs (e.g., the quantity of tortilla chips produced). These models are successful for their intended purpose. But they fail when they are used to predict the behavior of individuals cooperating in small groups (Smith, 2003). Is this another failure of human rationality?

8.1 A Puzzle from Behavioral Economics

Cooperation can be studied in the laboratory by having people interact in games in which the monetary payoffs for different choices are carefully controlled—dictator games, prisoner's dilemma games, bargaining games (e.g., the ultimatum game), trust/investment games, public goods games, and others. When behavioral economists used these methods to test predictions of game theory, they found that people in small groups do not act as if they are maximizing immediate monetary payoffs (e.g., Hoffman et al., 1998). In a one-shot interaction with anonymous others, *Homo economicus* models predict no generosity, no cooperation, no trust, and no punishment. Yet people give more, cooperate more, trust more, and punish defections more than these models predict, even when the experimenter tells them that the interaction is one-shot and anonymous. But why? Generosity in anonymous, one-shot games is irrational, given the standard theories. According to both economic *and* evolutionary game theory, repeated interactions are necessary for behaviors like this to evolve.

This "excess altruism" is considered irrational on many economic theories, and some social scientists have viewed it as evidence that the psychology of cooperation was shaped by group selection rather than selection operating on individuals (e.g., Bowles & Gintis, 2013). But are these behaviors really *excess* altruism—that is, beyond what can be explained by selection on individuals for direct reciprocity?

8.2 Adaptationist Game Theory: Model the Information-Processing Problem and Let the Psychology Evolve in Response

Selection does not occur in a vacuum: the physical and social ecology of a species shapes the design of its adaptations, and our hunter–gatherer ancestors lived in small, interdependent bands, and had many encounters with individuals from neighboring bands. Adaptations for direct reciprocity evolved to regulate cooperation in an ancestral world in which most interactions were repeated. The high prior probability that any given interaction will

be repeated should be reflected in their design. So should the fact that you have interacted at least once with a person: models of this social ecology show that meeting an individual once is itself a good cue that you will meet again (Krasnow, Delton, Tooby, & Cosmides, 2013).

Whether an interaction is one-shot or repeated is a judgment made under uncertainty. What decision rules does selection favor when this judgment-under-uncertainty problem is modeled?

8.3 The Evolution of Generosity under Uncertainty

In most agent-based simulations of the evolution of cooperation, the behavioral strategies are particulate—they do not have internal cognitive components that can evolve—and the simulation environment has either one-shot or repeated interactions, but not both. But what happens if these strategies have a cognitive architecture that can evolve, and the social environment includes both one-shot *and* repeated interactions, as in real life? It turns out that generosity in one-shot interactions evolves easily when natural selection shapes decision systems for regulating two-person reciprocity (exchange) under conditions of uncertainty (Delton, Krasnow, Cosmides, & Tooby, 2011).

In real life, you never know with certainty that you will interact with a person once and only once. Categorizing an interaction as one-shot or repeated is always a judgment made under uncertainty, based on probabilistic cues (e.g., Am I far from home? Did he marry into my band?). In deciding whether to initiate a cooperative relationship, a contingent cooperator must use these cues to make trade-offs between two different kinds of errors: (i) false positives, in which a one-shot interaction is mistakenly categorized as a repeated interaction, and (ii) misses, in which a repeated interaction is mistakenly categorized as one-shot. A miss is a missed opportunity to harvest gains in trade from a long string of mutually beneficial interactions. In a population of contingent cooperators, the cost of a miss is usually much higher than the cost of a false positive.

Using agent-based simulations, Delton et al. (2011) showed that, under a wide range of conditions, individual-level selection favors computational designs that decide to cooperate with new partners, even in a world where most of the interactions are one-shot. Each new partner comes with a number—a cue summary—that serves as a hint to whether an agent's interaction with that partner will be one-shot or repeated. The cue summaries are never perfect predictors: they are drawn from one of two normal distributions (one-shot vs. repeated) that overlap by different amounts.

In one set of simulations, agents evolve a decision threshold determining how strong cues that the interaction is one-shot must be before the agent defects. Selection favored a threshold of evidence so high that most interactions were classified as repeated, triggering cooperation. In other simulations, the agents were perfect Bayesians who developed rational beliefs about whether an interaction is one-shot by using Bayes' rule to integrate (i) cues that the given interaction is one-shot, with (ii) perfect knowledge of the base rate of one-shot interactions in the population. What evolved is a regulatory variable that determines the probability the agent will cooperate *given its rational belief that the interaction is one-shot*.

Selection favored designs with a very high probability (70%–90%) of cooperating given the rational belief that the interaction is one-shot, with modest gains in trade and a modest number of encounters for those interactions that were repeated. This was true even when the base rate of one-shot interactions was unrealistically high (50%–70%).

The simulations with Bayesian agents are particularly apt because most subjects who cooperate in experimental economics games say they believed the experimenter's claim (a cue!) that their interaction would be one-shot. The results show that natural selection can favor a disposition to start out cooperating, even in people who rationally believe an interaction is most likely to be one-shot. No group selection is needed. And this disposition to cooperate in one-shot interactions is not a mistake or an irrational "bias": it is the most adaptive, fitness-promoting decision.

References

Axelrod, R. (1984). *The evolution of cooperation*. New York; NY: Basic Books.

Bowles, S., & Gintis, H. (2013). *A cooperative species: Human reciprocity and its evolution*. Princeton, NJ: Princeton University Press.

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*, 325–370.

Cosmides, L., & Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (pp. 277–306). Cambridge, MA: MIT Press.

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

Cosmides, L., & Tooby, J. (2008). Can a general deontic logic capture the facts of human moral reasoning? How the

mind interprets social exchange rules and detects cheaters. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (pp. 53–119). Cambridge, MA: MIT Press.

Cosmides, L., & Tooby, J. (2015). Adaptations for reasoning about social exchange. In D. M. Buss (Ed.), *The handbook of evolutionary psychology: Vol. 2. Integrations* (2nd ed., pp. 625–668). Hoboken, NJ: Wiley.

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, 108*, 13335–13340.

Gigerenzer, G. (2014). *Risk savvy: How to make good decisions*. New York, NY: Viking.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychological Review*, *106*, 425–430.

Higham, J. P. (2014). How does honest costly signaling work? *Behavioral Ecology*, *25*(1), 8–11.

Hoffman, E., McCabe, K., & Smith, V. (1998). Behavioral foundations of reciprocity: Experimental economics and evolutionary psychology. *Economic Inquiry*, *36*, 335–352.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Macmillan.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*, 123–141.

Kleiter, G. D. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). New York, NY: Springer.

Krasnow, M. M., Delton, A. W., Tooby, J., & Cosmides, L. (2013). Meeting now suggests we will meet again: Implications for debates on the evolution of cooperation. *Nature Scientific Reports*, *3*, 1747.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (2nd ed., pp. 380–402). Oxford, England: Blackwell.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge, England: Cambridge University Press.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–111.

Mishra, S., & Fiddick, L. (2012). Beyond gains and losses: The effect of need on risky choice in framed decisions. *Journal of Personality and Social Psychology*, *102*(6), 1136–1147.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, *253*, 980–986.

Rips, L. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Rode, C., Cosmides, L., Hell, W., & Tooby, J. (1999). When and why do people avoid unknown probabilities in decisions under uncertainty? Testing some predictions from optimal foraging theory. *Cognition*, *72*, 269–304.

Smith, V. L. (2003). Constructivist and ecological rationality in economics (Nobel Prize Lecture, December 8, 2002). *American Economic Review*, *93*(3), 465–508.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39–67). Cambridge, England: Cambridge University Press.

Stephens, D. W., & Krebs, J. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York, NY: Oxford University Press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University 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