

MIT Press • Open Encyclopedia of Cognitive Science

Foundations of Rationality

Nick Chater¹

¹Warwick Business School

MIT Press

Published on: Sep 10, 2025

DOI: <https://doi.org/10.21428/e2759450.a7b4f718>

License: [Creative Commons Attribution-NonCommercial 4.0 International License \(CC-BY-NC 4.0\)](#).

Rationality is one of the most central concepts in cognitive science, yet also one of the least well defined. In some contexts, it is used interchangeably with terms like *intelligence* or even *reasonableness*. *Rationality* in its narrower conception is taken to be adherence to a set of normative standards that constitute justifiable reasoning. In both its broad and narrow senses, rationality is typically viewed as a prerequisite for interpreting the behavior of another person, artificial intelligence, or animal as having mental states at all. Attributing a belief, desire, or expectation to an entity implies an anticipation that it will act rationally given that belief, desire, or expectation—alongside the broader network of their other mental states. Some philosophers have gone further, arguing that rationality is essential to the very notion of the interpretability of language: that is, the idea that a person's utterances are not meaningless gibberish but instead follow, at least to some degree, from a justifiable pattern of reasoning. This perspective suggests that rationality underpins the human ability to reconstruct meaning and intention in communication.

History

From the late 19th century onwards, various mathematical methods have been developed to model different aspects of rationality. For example, various patterns of verbal reasoning have been formalized through increasingly sophisticated systems of *formal logic*. Indeed, it has often been supposed that the meaning of sentences can be captured in an underlying logical form (e.g., [Dowty et al., 1981](#)). This tradition, rooted in the work of the logicians and philosophers Gottlob Frege and Bertrand Russell more than a century ago, holds that logical inference determines the deductive implications of a sentence in virtue of its structure. So, for example, “All As are Bs” and “x is an A,” implies that x is a B, irrespective of the specific properties A or B or the identity of the particular object x. According to this viewpoint, one role of logic is to provide a mathematically precise way of capturing, and studying the properties of, deductive verbal arguments.

In parallel with the logical tradition, which aims to formalize inferences that can be made with deductive certainty, intuitive judgments about likelihood and uncertainty have been formalized using probability theory. Rational choice theory extends these ideas to decision-making, providing a formal framework for analyzing what counts as rational decision-making, whether in individual choices—balancing options and risks—or in strategic, interactive decisions, the domain of game theory. This framework has profoundly influenced economics, serving as a foundation for microeconomic analysis.

Despite these advances, formal models of rationality cover only narrow, well-defined domains. Everyday human reasoning—rich with metaphor, analogy, past experience, background knowledge, and countless other factors—remains largely resistant to formalization. Efforts to capture such reasoning in logic-based artificial intelligence frameworks during the 1970s and 1980s highlighted the complexity of the challenge but fell short of providing comprehensive models.

Core concepts

Styles of rational explanation

In cognitive science, many researchers have attempted to capture thought and behavior using formal methods such as logic, probability theory, decision theory, and game theory (e.g., [Griffiths et al., 2024](#)). This approach aligns with applications of rational choice theory in economics and the social sciences as well as behavioral ecology, in which animal behavior is often interpreted as a rational adaptation to environmental challenges.

Within this tradition are notable frameworks such as the notion of a computational-level explanation ([Marr, 1982](#)) and rational analysis ([Anderson, 1990](#)) [see [Rational Analysis](#)]. The common aim is to explain why human thought and behavior are so effective by demonstrating their adherence to rational standards, justified by what the agent knows or desires. Many theorists go further, arguing that the actual computations carried out by the mind follow, at least to some approximation, rational principles.

For instance, psychologists of reasoning proposed that the mind reasons verbally by carrying out inferential steps according to a *mental logic* ([Rips, 1994](#)). Similarly, many in Bayesian cognitive science argue that the brain's underlying representations and computations across domains—such as perception, categorization, causal reasoning, and learning—follow principles of rational probabilistic inference [see [Bayesian Models of Cognition](#)]. This perspective has been embodied most recently in the probabilistic language of thought framework [see [The Language of Thought Hypothesis](#)], which views cognition as underpinned by sophisticated rational calculations, using probabilistic inference over rich logical representations ([Piantadosi et al., 2016](#)).

Irrationality in the lab?

While many phenomena seem to be well explained by rational models, when human reasoning is tested directly in experiments, people appear to flagrantly and repeatedly violate formal rational principles. Thus, the psychology of verbal reasoning ([Evans et al., 1993](#)) and the vast and closely related fields of judgement and decision-making and behavioral economics ([Kahneman & Tversky, 2000](#)) have generated ever-growing catalogues of apparent reasoning blunders, inconsistencies, and biases that violate the principles of logic, probability, and rational choice theory.

This contrast has been explained in different ways. One approach is to postulate that the mind divides into two very different systems: an irrational system that is fast, intuitive, but error prone (*System 1*), and a more rational system, which is slow and deliberative (*System 2*; e.g., [Kahneman, 2011](#)). From a cognitive science perspective, this seems unsatisfying because rational cognitive models tend to be most successful in domains like vision and motor control, which seem to be paradigmatic *System 1* processes: they operate rapidly and without deliberation or reflection.

A second, and radical, viewpoint is that formal models of rationality should be abandoned entirely as descriptive models of thought (e.g., [Elqayam & Evans, 2011](#)). However, this approach runs the risk of failing to capture the obvious rational patterns in thought and behavior. For example, people formulate plans to achieve goals and follow them step by step, based on beliefs about the relevant states of the world. Indeed, as noted above, it has been widely argued that without some assumption of rationality, attributing mental states of any kind to agents at all seems unjustified.

A third approach is not to abandon formal models of rationality but to modify rational models to better account for how people actually think and reason: recognizing that rational thought is subject to severe computational limitations. For example, a natural proposal is that Bayesian probabilistic models are approximated by drawing samples from complex probability distributions rather than engaging in analytic probabilistic calculations. Indeed, this approach may systematically explain some patterns of apparent probabilistic reasoning errors ([Chater et al., 2020](#); [Dasgupta et al., 2017](#)). More broadly, the idea that cognition can be viewed as rational computation under resource constraints is a focus of current research ([Lieder & Griffiths, 2020](#)).

Questions, controversies, and new developments

Expanding the scope of formal theories of rationality

Alongside, and not necessarily in tension with, the restrictive approach to rationality is a more recent tradition in cognitive science that seeks to expand the scope of formal rational theories to incorporate the contextual and social richness of human interaction. This includes models of rational speech acts ([Frank & Goodman, 2012](#)), virtual bargaining in social interactions ([Chater et al., 2022](#)), and frameworks for reasoning about causality ([Pearl, 2000](#)), other minds ([Baker et al., 2017](#)), and morality ([Levine et al., 2024](#)). Interestingly, an important strand of this work reconceptualizes verbal reasoning as being primarily directed toward arguing and persuading rather than seeking objective truth ([Mercier & Sperber, 2011](#)). That is, verbal reasoning may be better understood as rhetoric in the traditional sense rather than as an exercise in pure reason [see [Reasoning and Argumentation](#)].

From this viewpoint, an important challenge for cognitive science is to develop formal models of the rationality underpinning the rhetorical games humans engage in. The rise of large language models with impressive verbal reasoning abilities presents a further challenge [see [Large Language Models](#)]. These models do not appear to be grounded in formal theories of reasoning. Rather, their capabilities emerge as a by-product of training deep neural networks on vast amounts of text. A key question for the future of cognitive science is whether such systems eliminate the need for formal theories of rationality in cognitive science or whether these models remain necessary to explain the operations of black-box artificial intelligence (AI) systems, just as they are required to understand the black-box neural network that is the human brain.

Rationality as emergent from social interaction

Another tantalizing possibility for cognitive science is that formal theories of rationality—and perhaps rationality more broadly—have been viewed from the wrong end of the telescope ([Mercier & Sperber, 2011](#)). Rather than treating rationality as primarily a property of individual minds and brains, it may be more appropriate to see at least some aspects of rationality as emerging from distributed computation arising from the interactions of many individuals operating over a shared cultural *workspace*—a public realm of argument, debate, observation and experimentation which allow humans collectively to refine ideas over time. Indeed, the most striking examples of rational systems in human culture—mathematics, science, legal reasoning, and financial and economic markets—seem less like the product of any single individual and more like vast edifices with complex (although imperfect) rational structure, which are continuously modified by many people over time ([Chater, 2015](#)).

From this perspective, perhaps formal theories of rationality—such as logic and probability—are not best viewed as in-built mechanisms of the human mind from which rationality arises. Rather, they may be the endpoints of extended cultural processes of reflection, negotiation, refinement, and experimentation [see [Cultural Evolution](#)]. That is, they emerge not from within the brain itself but from the long history of public deliberation, debate, bargaining, and reflection (e.g., [Oaksford, 2024](#)). Rationality may, from this point of view, be primarily social rather than a property of individual minds.

Broader connections

Formal rational theories extend far beyond cognitive science. These formal accounts have their origins in philosophy and mathematics but have applications throughout computer science and AI (where computations are typically seen as approximating some rationally justified computation), in the *rational actor* model that is dominant in modern economics, and extends further to many areas of social science and to animal behavior and in models of neural computation in the brain sciences. Indeed, any aspect of intelligent behavior—whether human, nonhuman animal, or AI—seems to demand an explanation of both why the behavior is intelligent and how such intelligence is achieved. Theories based on formal models of rationality attempt to answer these questions.

Acknowledgments

The author gratefully acknowledges support from the European Union/UKRI under Horizon Europe Programme Grant Agreement no. 101120763—TANGO, ESRC UKRI/NSF—Grant Ref: ES/Z504397/1 and Advanced Research and Invention Agency Collaboration Agreement SCNI-PR01-P16. The views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Further reading

- Briggs, R. A. (2023). Normative theories of rational choice: Expected utility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2023 Edition). Stanford University. <https://plato.stanford.edu/archives/win2023/entries/rationality-normative-utility/>
- Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (2024). *Bayesian models of cognition: The probabilistic approach to the mind*. MIT Press.
- Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, 71, 305–330. <https://doi.org/10.1146/annurev-psych-010419-051132>
- Wheeler, G. (2024). Bounded rationality. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2024 Edition). Stanford University. <https://plato.stanford.edu/entries/bounded-rationality/>

References

- Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates.
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing. *Nature Human Behaviour*, 1, 0064. <https://doi.org/10.1038/s41562-017-0064>
- Chater, N. (2015). Can cognitive science create a cognitive economics? *Cognition*, 135, 52–55. <https://doi.org/10.1016/j.cognition.2014.11.020>
- Chater, N., Zeitoun, H., & Melkonyan, T. A. (2022). The paradox of social interaction: Shared intentionality, we-reasoning, and virtual bargaining. *Psychological Review*, 129(3), 415–437. <https://doi.org/10.1037/rev0000343wrap.warwick.ac.uk+6>
- Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrà, P., & Sanborn, A. N. (2020). Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science*, 29(5), 506–512. <https://doi.org/10.1177/0963721420954801>
- Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, 96, 1–25. <https://doi.org/10.1016/j.cogpsych.2017.05.001>
- Dowty, D. R., Wall, R. E., & Peters, S. (1980). Introduction to Montague Semantics. In *Studies in Linguistics and Philosophy*. Springer Netherlands. <https://doi.org/10.1007/978-94-009-9065-4>

- Elqayam, S., & Evans, J. S. B. (2011). Subtracting “ought” from “is”: Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, 34(5), 233-248.

<https://doi.org/10.1017/S0140525X1100001X>

↩

- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Lawrence Erlbaum Associates.

↩

- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998. <https://doi.org/10.1126/science.1218633>

↩

- Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (2024). *Bayesian models of cognition: The probabilistic approach to the mind*. MIT Press.

↩

- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

↩

- Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge University Press.

↩

- Levine, S., Chater, N., Tenenbaum, J. B., & Cushman, F. (2024). Resource-rational contractualism: A triple theory of moral cognition. *Behavioral and Brain Sciences*. Advance online publication.

<https://doi.org/10.1017/S0140525X24001067>

↩

- Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, e1.

<https://doi.org/10.1017/S0140525X1900061X>

↩

- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company.

↩

- Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2), 57-74. <https://doi.org/10.1017/S0140525X10000968>

↩

- Oaksford, M. (2024). Social rationality and human reasoning: Logical expressivism and the flat mind. *Topics in Cognitive Science*. Advance online publication. <https://doi.org/10.1111/tops.12757>

[↩](#)

- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

[↩](#)

- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4), 392. <https://doi.org/10.1037/a0039980>

[↩](#)

- Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.

[↩](#)