Closing the Loop in Cognitive Science:

The Diachronic Evidential Strategy of Bounded Rational Analysis

Brendan Fleig-Goldstein

Department of Philosophy, Brown University, Providence, RI

brendan_fleig-goldstein@brown.edu

January 2026

Abstract

Why might a scientist want to establish a cognitive model as optimally suited to some particular environment? In this paper, I suggest that an unexamined motivation for establishing models as optimal is to uncover systematic discrepancies between idealized human behavior and observed human behavior. These discrepancies can lead to the discovery of previously unknown cognitive architecture details (e.g., resource constraints), which can then be incorporated into models and give rise to new idealized models that factor in these newly uncovered details. Further discrepancies then arise, and the process repeats itself in an iterative fashion. Most importantly, each step in this process provides evidence for descriptive claims about human cognition. The point, then, of establishing optimal models is to facilitate a particular process for marshalling evidence about the human cognitive system.

1 Introduction

Why might a scientist want to establish a cognitive process as optimal relative to a goal and an environment? There are typically three responses to this question. First, we may want to explain why some aspect of cognition is the way it is. If we can establish that some aspect of an organism's cognition is X, show that X is optimal, and show that that aspect of cognition is X *because* it is optimal, we will have provided a plausible teleological explanation of that aspect of cognition (Danks, 2008). Second, we may want to defend the claim that, as an empirical fact, the design of human cognition is very nearly optimal. Third, we may want to learn what human cognition is like, and we may think that a good strategy would be, methodologically, to start by considering the optimal cognitive model. These three kinds of answers correspond to what, in the context of adaptationism in biology, have been called explanatory, empirical, and methodological adaptationism, respectively (Godfrey-Smith, 2001).

I propose that there is a fourth role for optimality claims to play in cognitive science. One might call this form "evidential" in contrast to the other three forms. The basic idea is that the role of optimality claims is to allow for a specific strategy for marshalling evidence for descriptive facts about the cognitive system, which this paper will be devoted to laying out. In the original three forms, optimality claims do not play a role in providing evidential support for theoretical scientific claims. In the case of explanatory adaptationism, for example, optimality claims provide teleological explanation once a descriptive model of cognition has been evidenced; they are not thought to aid in the evidencing of the descriptive model in the first place. In empirical adaptationism, optimality claims are themselves hypotheses needing evidence (Orzack and Sober, 1994a). In methodological adaptationism, optimality serves as "no

more" than a *heuristic* guiding the search for models (Godfrey-Smith, 2001, p. 4). It does not help supply any specific answer to the question of how to test these models.

In this paper, my goal is to show how optimality claims are able to facilitate a particular testing strategy, and how this form of evidential reasoning is manifest in a cognitive science research program that has been variously called bounded rational analysis, resource rational analysis, and more (Icard, 2014; Griffiths et al., 2015; Lieder & Griffiths, 2020; Gershman et al., 2015; Icard, 2023). In Anderson's (1990) original version of rational analysis, scientists adopt an optimality assumption, characterize an agent's goal and environment, and then derive an optimal behavior function (i.e., a model of how the organism will be behave in response to different environmental inputs). That is, one supposes that an animal's behavior qua cognitive agent is very nearly optimally suited to achieving its—or its genes'—goals in some particular environment. If one models these three variables in some context and finds such an optimal relationship does not inhere, one attempts to revise these three variables until such a fit is obtained. If such a fit cannot be obtained, however, the principles of rational analysis in general give no prescription for how to proceed.

In bounded rational analysis, however, such a mismatch between the observed fit of these three variables and the optimal fit is the very beginning of inquiry. The driving assumption of bounded rational analysis is that the mind makes (near) optimal use of its limited cognitive resources in order to (near) optimally achieve its goals. Thus the fit of optimality is between goals, behavior, environment, and internal cognitive architecture. When the first three variables are found to deviate from optimality, bounded rational analysis provides principles for reasoning about sources of these discrepancies in the form of computational and metabolic resource costs, computational approximation and implementation schemes, cognitive parameters and other

functional details (Icard, 2018; Hébert & Woodford, 2023). Sources can be identified, incorporated into theory, and new discrepancies then emerge. An indefinitely iterative process can occur because optimality, in this picture, now holds as a relation between all the parts of the cognitive system, and not just between the organism's behavior and the environment.

In order to explain in-depth the nature of this logic of theory-testing, and consequently how optimality claims can play a role in the marshalling of evidence, this paper explores an extended analogy between Newtonian gravity research and bounded rational analysis. The logic of theory-testing concerned with here—in which discrepancies between theoretical prediction and observation serve to aid in the development and evidencing of theory in a systematic and iterative fashion—was first described by Smith (2014) in his analysis of the evidence attained in Newtonian gravity research. I argue that this analysis of evidence is appropriate for bounded rational analysis as well.

Toward the development of this point, this paper transitions back and forth between describing the challenges to, and the surmounting of these challenges by, Newtonian gravity research on the one hand, and the parallels in the case of cognitive science on the other. Bounded rational analysis is a specific variety of rational analysis (Anderson, 1990), which is in turn a specific variety of the "top-down" approach in cognitive science, in which scientists start from human behavioral data, infer cognitive processes, and successively develop more detailed mechanistic models of how such cognitive processes are implemented (as opposed to the bottom-up approach which starts with physiological data) (e.g., Marr, 1982). The first section of this paper is devoted to describing the nature of the scientific challenge facing gravity research and cognitive science. The second section describes how Newton and the top-down approach offer similar methodological solutions to these challenges, and how rational analysis extends the

logic of the top-down approach to become even more methodologically Newtonian. The third section describes how bounded rational analysis in turn extends the logic of rational analysis to finally enable research to realize Newtonian gravity research's iterative evidential logic. The fourth section describes in-depth an example from the bounded rational analysis literature to illustrate this evidential logic. The fifth and final section explores unresolved issues for thinking about evidence in this way.

Ultimately, the point of this paper is to offer a defense of the value of making optimality claims in cognitive science. The ever-popular Bayesian boom in cognitive modelling initiated by Tenenbaum and Griffiths (2001) began as heir to Anderson's rational analysis (Griffiths, Chater, and Tenenbaum, 2024). Yet Bayesian modelers have increasingly retreated from optimality claims—both voluntarily and under pressure from critics (e.g., Jones & Love, 2011; Bowers & Davis, 2012; Marcus & Davis, 2013; Crupi & Calzavarini, 2023). Many now wish to "keep the Bayesian baby but throw the optimality water out" (Frank, 2013) and offer purely descriptive models of cognition using the language of Bayesian modelling, with model parameters set through Bayesian data analysis (e.g., Tauber et al., 2017; Lee et al. 2018; Baribault & Collins, 2025). I believe that most of these conversations have presupposed either explanatory, empirical, or methodological flavors of optimality, and have overlooked the ability of optimality claims to play a role in the marshalling of evidence for descriptive models.

Many of those who wish to retain optimality considerations have moved to resource rational analysis. So far, however, there is a need to analyze why and how this approach works, and the nature of the evidence it produces, an analysis this paper develops.

2 The Obstacles to Developing Evidence in Celestial Mechanics and Cognitive Science

At the end of Smith (2014), Smith cautions against applying his analysis of evidence and knowledge attained in gravity research to other areas of scientific research, reminding the reader that only the careful examination of the history of evidence can bring forth an analysis for a particular field: "better we should identify the fundamental obstacles to developing evidence in each area of research and ask how they have been surmounted historically before we attempt any general account" (p. 343). Work has subsequently done so for other areas of science (e.g., Bokulich, 2020). Let me, therefore, briefly consider these obstacles faced by cognitive science and gravity research, before moving on to an analysis of the evidence in bounded rational analysis. Consider Smith's comment on the nature of the evidential problem:

> The inability to intervene, however, has not been the main evidential problem in gravity research on orbits. The one source of evidence, even after space flight, has been the orbital motions themselves, and these are extraordinarily complicated, opening the way to multiple representations of them to whatever level of precision is then current. Newton saw the possibility of turning those very complexities into a continuing source of evidence, provided the world turned out to be simple in one crucial respect, namely, that the motions are predominantly gravitational phenomena. There is, however, no reason to think that the main evidential problem in other areas of research lies in the complexity of the sources of evidence, and hence no reason to think that Newton's solution to the problem of evidence in gravity research applies elsewhere. In optics, for example, the problem is not the complexity of sources of evidence, but the complexity of light itself. (p 343)

That is, universal gravitation is not prohibitively complex; the motions of bodies that result from universal gravitation are prohibitively complex. In cognitive science, we want to

understand the principles and patterns that underlie an agent's perception, thought, and actions (i.e., behavior qua cognitive agent). Neural activity and human behavior are dauntingly intricate and present a challenge in and of itself to describe. But the hope is that there are principles and patterns that are less daunting. This may be for the reason that there is a level of abstraction away from neural activity that provides a bridge between biophysical mechanism and behavior. Since the cognitive revolution, the computer-mind analogy has been thought to provide a theoretically rich idea of what this intermediary-level of analysis is supposed to be like: namely, descriptions of informational states and their computational transformations.

There may also be organizing principles and patterns that are more than mere abstract descriptions of the neural hardware. For example, consider Dennett's (1971) design stance and intentional stance. These are strategies one can adopt for the purpose of explaining and predicting the behavior of different systems. The physical stance, design stance, and intentional stance each suppose that a system conforms to norms of, respectively: physics; good engineering principles; and intelligent or rational systems. The point of adopting these last two stances is exactly that trying to explain or predict the behavior of a system from the physical stance is often prohibitively expensive.

In either case, the problem of studying human cognition, it is plausible to think, does not lie in the complexity of the principles and patterns that underlie cognition, but in the sources of evidence for these principles and patterns—that is, the neural activity and human behavior. Minds may not be prohibitively complex, even if the behavior of a mind in a real world environment is.[1]

[1] It is worth noting one respect in which the challenge facing cognitive science and Newton differ: as Smith notes (2007), one of Newton's primary concern was to develop a method for discovering the true motions of the planets, in contrast to their apparent motion. It is not clear there is an appropriate analog for this problem in cognitive science.

The very fact that, as has already been mentioned, the top-down approach is interested in an intermediary-level between neural activity and behavior (as opposed to say, behaviorism or varieties of neuroscience that reject such a level of analysis) is a reason for thinking Newton's methodology is applicable. At the time of Newton, the then-dominant mechanical philosophy, as exemplified by Descartes and Huygens, sought to explain the motions of planets directly in terms of contact or impact interactions between substances and their geometric properties. Newton's goal, however, was to inquire into the forces at play in the solar system that gave rise to planetary motion, and notoriously abstracted away from the mechanism of the delivery of force. Most notably, Newton argued for a universal inverse-square attractive force between all particles of matter in the universe, but left unspecified how such action at a distance occurred. Newton was not giving mere mathematical descriptions of motions, or merely providing useful models for calculating planetary motion without concern for how these models relate to the solar system beyond their predictive success. Newton was inquiring into an intermediary-level between (contact) mechanism and motion, just as in cognitive science we are interested in an intermediary-level between neural activity and behavior.

The question now becomes how to inquire into such an intermediary-level of analysis. An attractive force between two physical bodies is not something that can be directly observed, but must be inferred from the motions of the two bodies. One strategy would be to hypothesize a reasonable model of forces at work in the solar system—one that goes well beyond the data—derive observable consequences from such a model, and test these against experience. In his 1670's Optical Lectures, Newton referred to this strategy as "the method of hypotheses." Huygens provides the clearest statement on the method of hypotheses as a strategy for scientific inference in his treatise on light (1690):

> Here principles are tested by the inferences which are derivable from them. The nature of the subject permits of no other treatment. It is possible, however, in this way to establish a probability which is little short of certainty. This is the case when the consequences of the assumed principles are in perfect accord with the observed phenomena, and especially when these verifications are numerous; but above all when one employs the hypothesis to predict new phenomena and finds his expectations realized. (p. 454)

One hypothesizes hidden structures of the world. Evidence comes when such a hypothesis leads to accurate prediction—especially when such predictions are novel and surprising. Newton famously rejected the method of hypotheses, as expressed in his General Scholium first added to the second addition of the *Principia*:

> I have not as yet been able to deduce from phenomena the reason for these properties of gravity, and I do not feign hypotheses. For whatever is not deduced from the phenomena must be called a hypothesis; and hypotheses, whether metaphysical or physical, or based on occult qualities, or mechanical, have no place in experimental philosophy. In this experimental philosophy, propositions are deduced from the phenomena and are made general by induction. (p. 943)

As an alternative to the method of hypotheses, Newton establishes generic mathematical relationships between forces and motions. These are generic in the sense that the relationships Newton concerns himself with are in no sense specifically about forces and motions in our solar system, but about what sort of attractive force in the abstract would give rise to what sort of motion in the abstract. For example, in book I of the *Principia*, Newton develops an account of motions of objects under various centripetal forces. Assuming the irrelevance of other forces,

and abstracting away the underlying causes of forces, Newton derives the motion of a body under a linear attractive centripetal force, inverse-square attractive force, as well as inverse-cubic attractive force. Establishing these relationships allowed Newton to take observations of motions and "deductively" infer the presence of forces. For example, Newton shows that if Kepler's area rule—"bodies sweep out equal areas in equal times and their orbits are stationary"—holds very nearly true, then the force holding such a body in motion is a very nearly centripetal attractive force. The antecedent of this conditional holds true from observation of planetary orbits, and so Newton deduces from this phenomenon the existence of a very nearly centripetal attractive force between the planets and sun.

In cognitive science, there is an analogous problem: the problem of attempting to study states and processes that are not directly observable—but must instead be inferred from behavior—has been with empirical psychology since the birth of cognitive science. As soon as mental processes cease to be identified with behavioral dispositions or brain processes, but instead are related to abstract machines and their computations, for example, psychology is confronted with the problem of having to investigate hidden structures.

The top-down approach, as canonically expressed by Marr (1982), is one way of getting at these states and transformations. One begins with behavioral descriptions and makes inferences about the goals of, and problem faced by, the cognitive system.[2] One then uses an analysis of the problem faced by the cognitive system to begin developing a model of how the mind actually solves that problem—successively establishing more detailed models until, potentially, one reaches a description of neural activity.[3] There is not an orthodoxy in the

[2] For example, Marr gives the example of considering the problem faced by a bird's wings when trying to achieve flight. Before giving any consideration to a wing's anatomy, one can consider the general aerodynamic challenge that any flying thing faces. Doing so can establish necessary conditions for the possibility of flight, and tell us what a bird's wings must be doing before we know how they do it.

[3] Marr divided his framework into three levels of analysis. This division is principled upon his commitment to a privileged "algorithmic and representational level." He believed that the cognitive system has internal states which

top-down approach for how one goes from higher-levels models—in the sense of more abstract models—to lower or more detailed models. One way is for higher-level models to simply rule out as impossible large classes of lower-level models, and thereby constrain the class of possible models to consider. The next step then is often resorting to "the method of hypotheses," whereby one hypothesizes a reasonable cognitive model that far exceeds the data, derives consequences from this model, and then tests these consequences against experience. In such a situation, the top-down approach contributes to the process of marshalling evidence for a model only by ruling out alternative models.

In 1990, Anderson laid out an approach he called rational analysis. The aim, according to Anderson, was to provide a methodology for studying and modelling cognitive behavior while avoiding the need to initially make speculations about underlying cognitive mechanisms and implementational details. Anderson was troubled by the unobservability of cognitive mechanisms and the difficulty this posed for generating and selecting between different models of cognition: "We pull out of an infinite grab bag of mechanisms, bizarre creations whose only justification is that they predict the phenomena in a class of experiments. These mechanisms are becoming increasingly complex, and we wind up simulating them and trying to understand their behavior just as we try to understand the human" (p. 8). Anderson believed that the answer to this problem—an answer he thought was latently present in Marr's approach—was to adopt an

act as vehicles of representation, and that information processing is accomplished by the manipulating of these data structures. Thus, for Marr, there is a fact of the matter about what representations are used and what rules or algorithms operate over them. Proponents of the top-down approach need not take on board this commitment—or indeed even the commitment that thinking is computation. One can instead imagine a very general hierarchy of abstraction of descriptions of cognitive processes, neutral to what kind of predicates feature into those descriptions. Descriptions at the highest level will consist of "black box" operations, and the goal of the top-down approach is to progressively break open black box operations into smaller and smaller black box operations. It may or may not be possible (in principle or in practice) to reach enough detail that the basic operations specified in the lowest description can be cached out in terms of the biophysical operations of the brain.

optimality assumption: begin by provisionally taking it to be the case that human intelligent behavior is optimal relative to the goals of the cognitive system and a normal environment.

Rational analysis is consequently concerned with establishing optimal relationships between environment, agent goals, and behavior. What behavior will best fulfill my goal in this particular environment? The optimality assumption in rational analysis is in this sense an assumption about human practical or instrumental rationality—an ability to act so as to obtain one's desires—and not an assumption about whether human thought conforms to certain norms of belief revision or cohesion.[4] "Goals," in this context means more than just a utility assignment of states of the world for an organism, but a specification of the task or problem that the cognitive system is attempting to accomplish—in Marr's sense, as discussed above.

The relationships between the triple of environment, goal, and behavior are generic in exactly the same sense that Newton's mathematical relationships between forces and motions are: the relations we are concerned with here between these three elements hold for any agents, and are not in any sense specific to human minds, or what we think human minds are like. By establishing such relationships, a cognitive scientist can then take observations of the statistics of the environment and human behavior, and make careful inferences about the task that the cognitive system is performing (or indeed, from any two elements of the triple to the third).[5]

Anderson and Newton both explicitly cast their work as an alternative to the method of hypotheses. In Newton's case, Newton was diverging from the approach typified by Descartes

[4] Note that research predicated on this sort of optimality assumption is also different from evolutionary psychology in several ways, including: being neutral to what extent an optimal behavior is due to phylogeny or ontogeny; caring more about the statistics of a normal environment than about a specific evolutionary history; and being neutral to just about every contentious debate in evolutionary biology, such as varieties of adaptationism regarding evolutionary biology, whether evolution is gene-centered, etc.

[5] Cognitive tasks quite often require making uncertain inferences in stochastic environments. As such, cognitive behavior can rarely be shown to be optimal via a mathematical proof. Instead, simulations of computational agents in an environment with defined utility and cost functions often constitute the best means of establishing an agent as optimal.

and Huygens, in which contact mechanism models were offered, and mathematical consequences derived and tested against experience. In Anderson's case, Anderson was diverging from cognitive psychologists who were offering cognitive models "pulled from an infinite grab bag" and whose "only justification is that they predict the phenomena in a class of experiments." He instead seeks, like Newton, a non-arbitrary method of deducing from an observable phenomenon an unobservable state of affairs: in Anderson's case, a way of deducing from behavior and environment, facts about human psychology. All that is needed is the assumption that human psychology is optimal in the sense already described.[6] The following quote is from the end of Book 1, section 11 of the *Principia* and further illustrates the connection Newton's method has to the normative top-down approach.

> Mathematics requires an investigation of those quantities of forces and their proportions that follow from any conditions that may be supposed. Then, coming down to physics, these proportions must be compared with the phenomena, so that it may be found out which conditions of forces apply to each kind of attracting bodies. And then, finally, it will be possible to argue more securely concerning the physical species, physical causes, and physical proportions of these forces. (p. 588f)

3 The Research Strategy in Newtonian Gravity Research and Bounded Rational Analysis

I have stated now the sense in which Newton's method involves making "deductions from phenomena" as opposed to putting forth hypotheses and testing their derivable consequences, and how Anderson's method follows a similar strategy. But this is just the start of a complicated logic of theory-testing that Newton sets in motion for gravity research.

[6] As Smith points out, Newton's deductions from phenomena—e.g., conclusions about the sort of attractive forces present in our solar system, derived from motions—nevertheless require fallible assumptions. For example, Newton's deduction of very nearly inverse-square attraction assumes that the stars do not have a noticeable effect on orbital motion.

Critical to this process is the notion of what Smith calls Newtonian idealizations, which are licensed by Newton's fourth rule of reasoning in the *Principia*:

> In experimental philosophy, propositions gathered from phenomena by induction should be taken to be either exactly or very nearly true notwithstanding any contrary hypotheses, until yet other phenomena make such propositions either more exact or liable to exceptions. (p. 796)

Newton deduces very nearly inverse-square centripetal force between the planets and the sun from observed motions. Inexact inverse-square attraction is then made general by induction so as to hold between all particles of matter in the universe, and then "should be taken to be" exact by the fourth rule of reasoning—the final result being Newtonian universal gravitation.

The next step for gravity research is to then examine the discrepancy between calculated orbital motion (calculations made from universal gravitation) and observed motion. This discrepancy becomes what Smith calls a "second-order phenomenon"—called such because the discrepancy is not a feature of the world, but exists as a relationship between theoretical calculation and observation. The move from inverse-square attraction between the planets and sun to universal gravity yields a prescription for how to deal with such discrepancies. Universal gravitation entails that orbital motions *would* be exactly elliptical *were* there no unaccounted for forces affecting these orbits. Thus, when orbits fail to be perfectly elliptical, the next move is to identify the physical source of the discrepancy—for example, an undiscovered body of mass in the solar system exerting an unaccounted for attractive force on the other orbital bodies. Once this source has been identified and integrated into theory, theoretical prediction and observation achieve a greater agreement. Once theory has been adjusted, a new discrepancy between theory and observation emerges, and the above process can repeat itself. Universal gravitation therefore

facilitates a process of iteratively uncovering what Smith calls "differences that make a difference" by providing specifications of what sorts of physical details can be making a difference and exactly what kind of difference they would make.

The clearest example of this process comes from the discovery of the planet Neptune in 1846—a discovery predicted from the systematic discrepancies in the orbit of Uranus. Another example—and as Smith points out, a much more typical example in the history of gravity research—came directly before this episode. The discovery of a previously unaccounted for higher-order interaction effect between Jupiter and Saturn led to the resolution of a long-standing anomaly in the orbits of these two planets. The incorporation of this source of discrepancy into physical theory then uncovered the discrepancy in Uranus' orbit that led to the discovery of Neptune. These two examples that Smith uses illustrate the fact that physical sources can obscure or mask discrepancies caused by other physical sources—in this case, the unaccounted for higher-order interaction effect between Jupiter and Saturn masked the discrepancy with a clear signature in Uranus' orbit.

Each time a discrepancy between prediction and observation emerged in the history of gravity research, there were two possibilities: either universal gravitation was in fact wrong in some way, or else there were as yet unaccounted for physical sources of the discrepancies. When only supposing an inverse-square attractive force between planetary bodies and the sun, there is little sense in talking of unidentified physical sources of discrepancies. All that can be said is that the attraction between the planets and sun is not quite inverse-square. On the other hand, the law of universal gravitation and the laws of motion specify a range of physical details that can be making important differences, and specify exactly how these details will make a difference.

Now consider the move from rational analysis to bounded rational analysis. The former seeks only to look for the right connection between goals, environment, and behavior. Anderson believed that serious considerations of cognitive resource constraints (computational, biological, etc.) during rational analysis—because these attributes of a system are equally as unobservable as goals of a system—would add too many degrees of freedom and lead back to the method of hypotheses. Anderson's solution was to stipulate that only the "minimum" constraints should be assumed in a rational analysis, where "minimum" means considering only those constraints that must be true for all real-world cognitive agents and are *ipso facto* not speculative. For example, if a computation would require an embodied agent to use more space and time than is available in the universe, then that computation is intractable and can be excluded from consideration in a rational analysis.

Bounded rational analysis, instead of making the minimum amount of assumptions about resource constraints, makes such resource constraints full-fledged elements of consideration in a specification of optimal human behavior. Factoring in constraints, or bounds, in an analysis of cognition is not new. Herbert Simon's (1955) notion of bounded rationality is exactly the idea that human cognition is shaped by both rationality and bounds. For Simon, however, bounds were not a consideration in characterizing what it meant to be rational, but were a separate additional factor shaping human thought. Human thought is rational, Simon believed, until it runs out of computational space and time resources and stops: there is no design assurance that the cognitive strategy chosen makes the best use of that limited space and time. This latter idea is exactly what bounded rational analysis does take to be true. Optimality becomes a relationship between goals, behavior, environment, and resource constraints: humans make optimal use of their available biophysical and computational power in the pursuit of behaving optimally.

This claim, on one hand, doubles down on human optimality—optimality now holds "universally" between all elements of the cognitive system. On the other hand, such a position is a mere extension of the logic already contained in rational analysis. As Icard (2025) points out, bounded rational analysis amounts to simply recognizing that an organism's organs for thinking are a relevant part of the environment to which an organism's cognition is adapted. Nevertheless, Anderson is still right that this move adds further unknowns and threatens to undermine his original motivation of avoiding dangerous speculation. Icard (2025), echoing Anderson's comment that rational analysis is a high-risk high-gain enterprise, suggests that bounded rational analysis is a "higher-risk higher-gain enterprise."

The most pertinent question facing rational analysis is the question of why it makes sense to look for optimal models when we know that cognition deviates from optimality in clearly observable respects (Tversky & Kahneman, 1973; Marcus & Davis, 2013). The answer, I believe, is contained in Newton's rules of reasoning. One gives human research participants a cognitive task—for example, give them small samples and ask them to generalize to hypotheses (Tenenbaum & Griffiths, 2001).[7] Establish an optimal model either through running simulations or through a more direct proof method. Observe that human behavior is very nearly the same as the behavior of the optimal model; one has now deduced that human behavior in this context is very nearly optimal. Make this optimality claim general by induction (that is, that optimality

[7] This paper began the contemporary Bayesian boom. This paper's stated goal is generalizing Roger Shepard's 1987 paper "Toward a Universal Law of Generalization for Psychological Science." In that paper, Shepard writes: "The tercentenary of the publication, in 1687, of Newton's *Principia* prompts the question of whether psychological science has any hope of achieving a law that is comparable in generality (if not in predictive accuracy) to Newton's universal law of gravitation." Shepard called his psychological law "universal" because it is supposed to apply to all cognitive agents, not just human agents. Despite the fact that his use of the word "universal" is disanalogous to the sense in which gravitation is universal, Shepard is still capturing something deeply Newtonian. To arrive at his law, first, he derives a general principle the justification of which is not based on the structure of any organism or mind, but on the structure of the environment: namely, the structure of objects considered in a certain degree of abstraction. He then argues that we should expect any organism to approximately obey this general principle, by making explicit appeal to natural selection's tendency to endow organisms with adaptive features. Shepard is therefore engaging in a form of rational analysis, before Anderson coined the name. Or perhaps more accurately, Shepard is following the normative top-down approach.

holds very nearly for other cognitive tasks and at other levels of design—e.g., resource allocation). And by Newton's fourth rule of reasoning one *takes it to be exact*.

Now deviations from optimality become second-order phenomena themselves in need of explanation. Just as in the case of universal gravity, "universal optimality" specifies a range of cognitive details that can be making an important difference to the cognitive system, and exactly how these details will make a difference. The goal of research transforms into an attempt to establish ideal models for the express purpose of giving rise to discrepancies between optimal and human behavior. The hope is that there will be a clear signature, and that cognitive sources of these discrepancies can be identified and incorporated into theory yielding ever closer agreement between calculation and observation—in an iterative process that each time presupposes all previously identified cognitive sources as well as the principle of universal optimality itself.

4 The Example of Categorization

In this section, I consider the example of categorization to illustrate the way in which discrepancies between theory and observation can iteratively uncover cognitive differences that make a difference. Anderson's (1990) rational analysis of categorization began with a characterization of the cognitive task. The task of categorization, Anderson argued, is a special case of inferring an unknown property of an object from known properties—the special case being that the unknown property of interest is a category label. For example, one might encounter various kinds of insects with varying properties, including the occasional ability to sting you. When encountering new insects, one wants to infer from, for example, visual properties whether the insect is of the stinging variety. Anderson argued that the ideal way to accomplish this task would be to start by imagining every possible way to group the insects so

far encountered into stinging and non-stinging groups. Then one should derive likelihoods of seeing various visual features on a stinging or non-stinging insect given the different groupings (e.g., how many stinging insects in this grouping are black and yellow?). Finally, which insects actually are stinging or not-stinging allows one to update the probabilities assigned to each grouping (each hypothesis) using Bayesian inference.

Anderson acknowledged that the ideal solution—of calculating the most likely grouping by considering every possible grouping after every observation—is most certainly computationally intractable for realistic categorization problems. Anderson proposed that instead of considering every possible grouping at every encounter of a new object, humans approximate this ideal by "locking-in" the most likely grouping after each encounter, and only consider different ways of extending the grouping to include the next object.

Sanborn et al. (2006, 2010) criticized Anderson's analysis, arguing that this choice of modification to the ideal is arbitrary. As Griffiths et al. (2015) write, "building in this constraint [that humans lock in the most probable grouping at each stage] misses two opportunities: to compare human behavior to the ideal predictions from an unconstrained computational-level account, and to explore the consequences of adopting different approximation schemes" (p. 221). This assessment gets it exactly right. The provision in rational analysis for making the minimal assumptions about resource constraints fails to provide a principled means of moving from computationally unrealistic to realistic models of cognition. The result then is a recourse to the method of hypotheses, in which arbitrary modelling assumptions enter into theorizing.

Sanborn et al. (2006, 2010) instead investigated a range of ways to approximate Anderson's computationally intractable ideal model of categorization, including using Markov chain Monte Carlo and particle filter techniques. The challenge is to estimate the probability

distribution over the possible groupings of objects. A particle filter model, after observing a new object, instead of calculating the updated probabilities of every possible grouping, stochastically chooses a finite number of groupings to update and discards the rest. The quantity of "particles" in the model determines how many groupings are chosen at each stage to update, and groupings are chosen with the same probability as their currently assigned probability of being the correct grouping.

Sanborn et al. showed that Anderson's modified model was a special case of a particle filter model, in which a single particle deterministically chooses the current most likely grouping at each stage. Further, they showed that 100 particles could fairly accurately estimate the intractable ideal solution, and that a single (not deterministically drawn) particle filter model most accurately matches observed human performance.

The most convincing aspect of Sanborn et al.'s model is not in the predictive accuracy, nor in its ability to predict previously unobserved and surprising data, but in the fact that it is able to recreate very specific deviations from an ideal, and ultimately offer a plausible construal of the source of these discrepancies. Their single stochastic particle model recreated two major ways in which humans deviate from ideal categorization behavior—namely, the exhibition of order effects and posterior matching.

Order effects occur when human responses differ depending on the order of presentation of stimuli. While not in and of itself a second-order phenomenon, when it is recognized that order effects for a particular task are a deviation from an ideal, and therefore in need of explanation, order effects become a second-order phenomenon.

Posterior matching occurs when humans select a hypothesis in an inference problem with the same probability that an ideal model assigns to that hypothesis being true (Danks &

Eberhardt, 2011). For example, if the ideal categorization model assigns a 90% probability to a particular grouping, 90% of research participants will select that grouping as the true grouping. Similarly, however, and far more concerning, is the fact that if the ideal model assigns a 10% probability to a particular grouping, 10% of research participants will select that grouping.

Both of these effects are systematic deviations from optimal performance with a clear signature, which could not otherwise be observed without the establishment of an ideal model. Sanborn et al.'s model offers a plausible identification of the cognitive source of these discrepancies: namely, that humans approximate the ideal solution using a particle filter strategy.

To what extent, however, can we say that a single stochastic particle filter strategy is the optimal way to approximate the ideal solution? Icard (2014) explored this question by running simulations with different agents, including different particle filter agents, the single deterministic particle filter model corresponding to Anderson's model, and a couple others. In the first simulations, no assumptions about computational costs were made, and he found that the particle filter agent using 10 particles outperformed other agents using fewer particles.

By doing this simulation, Icard has generated a new second-order phenomenon: the discrepancy between the performance of the 10 particle filter model and the single particle filter model that more closely matches with human performance. The speculative answer Icard offers is that the resource cost of running 10 particles as opposed to a single particle outweighs the worth of the very slight improvement in performance:

> Simply looking at the marginal increase in (estimated) fitness of keeping two particles instead of one, we see that it would only be worth the extra time, space, and energy if such costs amount to less than 1-2% of a utile (or more generally, 1-2% of the difference between payoff with a correct and with an incorrect prediction). (p. 87)

Icard also found that the deterministic single particle filter model (called the MAP algorithm in the quotation below) outperforms slightly the stochastic single particle filter model that more closely matches human performance. Here again we have a discrepancy between human performance and a potentially more optimal agent. Icard reasons as follows:

> As reviewed in some detail in the previous chapter, there is good empirical evidence that computations in the brain are essentially noisy, and that we should view sampling algorithms such as the particle filter as *harnessing* [his emphasis] this noise to the agent's advantage, rather than adding noise to otherwise deterministic computations. From this perspective, it would require further energy and resources to eliminate this noise at each step, to bring us from the particle filter to the MAP algorithm. If this is the right way to think about it, and if these simulations are indicative of typical scenarios, then the cost of reducing noise would have to be less than 2-3% of a utile for it to be worth the effort. (p. 88)

In both cases, Icard provisionally takes it to be the case that humans are making optimal use of their limited cognitive resources, and uses this assumption to identify potential cognitive sources of the discrepancies that emerge in the iterative process of comparing optimal agents to human performance. The next step in Newton's methodology would be to take these identified sources of discrepancies—the lower bound on particle costs and the lower bound on the cost of reducing noise in computations—and incorporate these elements into theory. The result should be improved predictive accuracy in areas beyond just the specific discrepancies that licensed us to identify these sources in the first place. This last requirement is to protect against the possibility that we have introduced sources in an *ad hoc* fashion. Such a next step is a challenge for future research.

This example is one of many. Bounded rational analyses of numerical estimation tasks, for example, offer a similar iterative case study. Lieder, Griffiths, et al. (2018) show that anchoring bias—a seemingly irrational systematic deviation from optimal estimation—can be explained as the signature of resource-rational approximation using early termination of sampling (Lieder et al., 2012; Vul et al., 2014). Once this cognitive source is identified and incorporated into theory, the model is able to successfully reproduce the specific conditions under which anchoring effects will be stronger or weaker, including the "incentive anomaly" where financial rewards reduce bias for self-generated but not experimenter-provided anchors (Epley & Gilovich, 2005, 2006). Additional examples of using apparent deviations from optimality to reason about and subsequently incorporate constraints include work on memory (Gershman et al., 2015; Azeredo da Silveira et al., 2024), perception (Albert et al., 2012), metacognition (Lieder, Shenhav, et al., 2018), and attention (Hébert and Woodford, 2023). Lewis et al. (2014) provide a general mathematical framework, proving that various cognitive strategies are optimal if and only if specific utility functions, resource bounds, and environmental structures obtain—thereby formalizing the space of possible cognitive sources that can be systematically identified through this evidential process.

5 The Nature of the Evidence

In Smith's analysis, evidence enters into this process in three places. The first is when a discrepancy with a clear signature emerges after comparison of calculation and observation. The second is when a source of this discrepancy is identified. And the third is in the incorporation of this discrepancy, and the subsequent iteration of this process, each time presupposing all previously identified sources. I have shown, in the case of categorization, two rounds of such an

iteration process. First, when Sanborn et al. identified a particle filter strategy as a plausible source of the discrepancy between human categorization behavior and ideal performance—a discrepancy with a clear signature in the form of posterior matching and order effects. And second, when Icard identified specific cost bounds as plausible sources of the discrepancy between the number of particles humans appear to use, and the number of particles that would be optimal without such costs.

A few points on the identification of sources of discrepancies. The source of a discrepancy will almost always be massively underdetermined by the discrepancy. This is true for both bounded rational analysis and Newtonian gravity research. To use Smith's example again, there are many different masses and locations Neptune could have had that would have explained the deviations observed in the orbit of Uranus. The source of the discrepancy could have turned out to have not been another planet at all. In the same way, there will generally always be a range of cognitive sources that can equally well account for a discrepancy, even with a clear signature.

Looking at *table 1* (found at the end of this paper), it is possible to begin to catalog different kinds of sources of discrepancies. The left half of this table is taken from Smith (2014). I think the right half of this table will need to be modified as bounded rational analysis develops. The sources in blue are the most interesting: they tell us something totally new about the cognitive system. For example, Sanborn et al.'s work showed that a different resource-optimal approximation scheme was needed. Icard's work showed that there were resource costs not taken into account.

But discrepancies can also be due—not to unknown factors of the cognitive system, but—for example, to an insufficiently approximated (intractable) ideal solution. That is, human

performance might be closer to the ideal than we suspected, but we as scientists incorrectly estimated what the ideal is. Such a situation might occur because we did not run enough simulations, take enough samples in our approximations, or because we failed to consider alternative agents in our simulations—and as a consequence we have incorrectly assessed what constitutes ideal performance. This kind of source of discrepancy parallels the one in Newtonian gravity research that arises out of the difficulty of the three body problem. It has, historically, been difficult to mathematically determine the exact motions of planets predicted by Newtonian gravity, even given a perfect knowledge of all relevant locations and mass distributions. This possible source of discrepancy is referred to in the table as "insufficiently converged infinite-series calculations."

Another possible source of discrepancy is that the resource-optimal solution to a problem is not an approximation of the (unbounded) ideal computational solution, but instead an alternative computation altogether.[8] Sanborn et al. (2006; 2010), for example, only consider methods of approximating the ideal computation. Griffiths et al. (2015) is also a notable project in which rational analysis is extended to include considerations of computational costs in the way we have been thinking about. Here, too, the authors simply stipulate that the next step after establishing an ideal (Bayesian) model is to look at algorithms that approximate these computations and find the approximations scheme that most closely matches human performance. As Icard (2014) has noted, one goal of bounded rational analysis research should be to specify the conditions in which approximating the ideal computation really is the most

[8] Such a distinction is not easily made. At what point does a computation cease to be an approximation of a computation and start to be a totally different one? Might we say that a simple lookup table agent "approximates" a Bayesian ideal computation in the same way that a particle filter agent does? See Maloney and Mamassian (2009) for a proposal of how to behaviorally test whether an agent is implementing a lookup table method versus a Bayesian calculation scheme in which priors, likelihoods, prediction error and so forth are separately represented. The basic idea is to see whether learning transfers and generalizes in particular ways. These sorts of tests can help operationalize the distinction concerned with here.

resource-optimal solution. Icard (2017), for example, explores "broad conditions" in which approximately Bayesian agents would outperform other agents after factoring in computational costs. In this paper, Icard characterizes a plausible model for how agents might incur computational costs, and shows that agents of this sort would be optimal when approximating Bayesian computations.[9] Establishing these conditions allows for the potential identification of a source of discrepancies between calculation and observation: humans either are or are not agents of this sort.

Finally, it should be noted that evidence for the existence of cognitive sources of discrepancies can come in many forms. Consider again Icard's (2014) appeal to the notion that computations in the brain harness the noisiness of neural activity, and that therefore there is a cost to reducing noise during computation. Icard appeals to other research unrelated to bounded rational analysis to support this claim. Consider also how to test whether humans use the computational cost architecture just mentioned in the previous paragraph. The best way may simply be hypothetico-deductive: deriving surprising, novel, and accurate predictions and testing them against experience. As Smith notes, the fact that the testing of the existence of particular physical sources of discrepancies in gravity research often took a hypothetico-deductive form obscures the more general logic of theory-testing. The fact that hypothetico-deductive evidential reasoning retains a role in bounded rational analysis similarly has the potential to obscure the role of optimality claims in cognitive models.

6 Conclusion

It is more subjectively probable that two students cheated on an exam if they gave the same wrong answer than if they had given the same right answer. Exceedingly different

[9] This characterization draws on thermodynamics and information theory to abstractly characterize cost of computation.

cognitive models can achieve the same behavioral result. We should expect optimal agents to agree behaviorally. Thus, when an idealized model agrees with human performance, there is little reason to think that the algorithms used by the mind and the model are the same. Agreement between model and human performance—to the degree that that performance is optimal—is always explained away. The remainder is the deviation from optimality, and this remainder constitutes one of the best sources of evidence we have available in cognitive science. This might be called the *Anna Karenina* Principle, called such after Leo Tolstoy's opening lines: "Happy families are all alike; every unhappy family is unhappy in its own way."

Deviations from optimality can only be observed by establishing optimal models. The goal of bounded rational analysis is to establish optimal models, and to iteratively factor in limitations of the system so as to be able to make new optimal models and continue this evidential process long after "full-blown" optimality has been left behind. Thus optimality claims play a role in the process of testing theories of cognition, and thereby go beyond either empirical, methodological, or explanatory roles.

There are many reasons one might be averse to optimality claims. Herbert Simon (1991), in response to Anderson's proposal of rational analysis, wrote that "our interest is in the learning process itself—not a hypothetical one or an optimal one, but the one that people use" (p. 35). There has been much call within the program of Bayesian models of cognition to set optimality aside (Frank, 2013). These calls acknowledge that humans are likely not optimal in all contexts, and consequently optimality will not always be particularly methodologically useful.[10] And of course, the debate concerning the degree to which human thought is or is not "rational," stretches

[10] Tauber et al. (2017) instead suggest that we set parameters in a Bayesian model of cognition through Bayesian data analysis. In other words, they suggest we make design choices for our models based on how well different models predict the data. If my analysis of the evidence in bounded rational analysis is correct, however, we need not be Bayesian—re scientific evidence—about our Bayesian models of cognition. The logic of theory-testing, instead, can match one closer to that of Newtonian gravity research.

back decades, with a strong school of thought initiated by the work of Tversky and Kahneman (1973) decidedly arguing that it is not.

I propose that, while it might seem that there are two views of human cognitive design—one that says that human cognition is well adapted to thinking and one that says it is less than well adapted—this is not a question that can be determined at present. Even if we have good reason to believe that we are not making the best use of our limited resources, we will need to uncover all the differences that make a difference to cognition in order to be able to say exactly in what ways that is the case.

Smith notes that what survived the transition from Newtonian to Einsteinian gravity were all the details that made a difference—the presence and distribution of mass throughout the solar system, their motions, and so on. These details, in this sense, have the best claim to knowledge attained in the course of Newtonian gravity research. In the same way, after researchers can declare exactly the ways in which the assumptions of bounded rational analysis fail, the details that make a difference—the cognitive parameters, the specification of computational architecture, the metabolic costs, and so forth—these will remain in place and will hopefully have the best claim to knowledge attained in cognitive science.

References

Albert, M. V., Catz, N., Thier, P., & Kording, K. (2012). Saccadic gain adaptation is predicted by the statistics of natural fluctuations in oculomotor function. Frontiers in Computational Neuroscience, 6, 96.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Anderson, J. R. (1991). Is human cognition adaptive?. *Behavioral and Brain Sciences*, *14*(3), 471-485.

Azeredo da Silveira, R., Sung, Y., & Woodford, M. (2024). Optimally imprecise memory and biased forecasts. American Economic Review, 114(10), 3075–3118.

Baribault, B., & Collins, A. G. (2025). Troubleshooting Bayesian cognitive models. *Psychological Methods*, *30*(1), 128.

Bokulich, A. (2020). Calibration, coherence, and consilience in radiometric measures of geologic time. *Philosophy of Science*, *87*(3), 425-456.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, *138*(3), 389.

Bowers, J. S., & Davis, C. J. (2012). Is that what Bayesians believe? reply to Griffiths, Chater, Norris, and Pouget (2012).

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, *3*(2), 57-65.

Crupi, V., & Calzavarini, F. (2023). Critique of pure Bayesian cognitive science: A view from the philosophy of science. *European Journal for Philosophy of Science*, *13*(3), 28.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. *The probabilistic mind: Prospects for Bayesian cognitive science*, 59-75.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87-106.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines*, *21*(3), 389-410.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. Journal of behavioral decision making, 18(3), 199–212.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. Psychological science, 17(4), 311–318.

Frank, M. C. (2013). Throwing out the bayesian baby with the optimal bathwater: Response to. Cognition, 128(3), 417-423.

Gershman, S. J. (2021). The rational analysis of memory. In J. H. Byrne (Ed.), The oxford handbook of human memory. Oxford University Press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349, 273–278. https://doi.org/10.1126/science.aac6076

Goodman, N. D., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P. W., & Hamrick, J. B. (2015). Relevant and robust: A response to Marcus and Davis (2013). *Psychological science*, *26*(4), 539-541.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, *1*(1), 107-143.

Godfrey-Smith, P. (2001). Three kinds of adaptationism. *Adaptationism and optimality*, 335-357.

Goodman, N. D., & Tenenbaum, J. B. (2014). Probabilistic models of cognition. Online, http://probmods.org.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012).

Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (Eds.). (2024). *Bayesian models of cognition: Reverse engineering the mind*. MIT Press.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217-229.

Hébert, B., & Woodford, M. (2023). Rational inattention when decisions take time. Journal of Economic Theory, 208, 105612.

Huygens, C. (1897). *Huygens Christiaan: oeuvres complètes*(Vol. 1). M. Nijhoff.

Icard, T. (2014). The Algorithmic Mind.

Icard, T. (2014, January). Toward boundedly rational analysis. In *Proceedings of the Cognitive Science Society* (Vol. 36, No. 36).

Icard, T. (2017). Bayes, Bounds, and Rational Analysis. *Philosophy of Science*.

Icard, T. (2025). Resource Rationality. *Manuscript*.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169-188.

Körding, K. (2025, March). The simple conceptual logic of normative models of behavior [Substack blog post]. https://kording.substack.com/p/the-simple-conceptual-logic-of-normative

Lee, M. D. (2018). Bayesian methods in cognitive modeling. *The Stevens' handbook of experimental psychology and cognitive neuroscience*, *5*, 37-84.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. Topics in Cognitive Science, 6, 279–311. https://doi.org/10.1111/tops.12086

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.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. Psychonomic bulletin & review, 25(1), 322–349.

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. Advances in neural information processing systems, 25.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. PLoS computational biology, 14(4), e1006043.

Ma, W. J., Kording, K. P., & Goldreich, D. (2023). Bayesian models of perception and action: An introduction. MIT press.

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: testing Bayesian transfer. *Visual neuroscience*, *26*(1), 147-155.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition?. *Psychological science*, *24*(12), 2351-2360.

Marr, D. (1982). Vision. W.H. Freeman and Company.

Newton, I., Cohen, I. B., & Whitman, A. (1999). The Principia, mathematical principles of natural philosophy, a new translation by I. *Bernard Cohen and Anne Whitman (University of California, Berkeley, 1999)*.

Orzack, S. H., & Sober, E. (1994). Optimality models and the test of adaptationism. The American Naturalist, 143(3), 361-380.

Orzack, S. H., & Forber, P. (2010). Adaptationism. *Stanford Encyclopedia of Philosophy*.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99-118.

Simon, H. A. (1991). Cognitive architectures and rational analysis: Comment. In *Architectures for intelligence: The 22nd carnegie mellon symposium on cognition* (pp. 25-39).

Smith, G. E. (2004). The methodology of the Principia. In *The Cambridge companion to Newton*, ed. I. Bernard Cohen, and George E. Smith, 138–173. Cambridge: Cambridge University Press.

Smith, G. E. (2007). Newton's Philosophiae Naturalis Principia Mathematica. *Stanford Encyclopedia of Philosophy*.

Smith, G. E. (2012). How Newton's Principia changed physics. In *Interpreting Newton. Critical essays*, ed. Andrew Janiak, and Eric Schliesser, 360–395. Cambridge: Cambridge University Press.

Smith, G. E. (2014). Closing the Loop: Testing Newtonian Gravity, Then and Now. In Newton and Empiricism, ed. Zvi Beiner, and Eric Schliesser, 262-351. Oxford: Oxford University Press.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. Psychological Review, 124(4), 410.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. Behavioral and brain sciences, 24(4), 629-640.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. Cognitive science, 38(4), 599–637.

Table 1

"The Many Sources of Discrepancies"

Newtonian gravity research (taken from Smith, 2014)		Bounded rational analysis	
In observations	In theoretical calculations	In observations	In theoretical calculations
1. Simple error - "bad data"	1. Undetected calculation errors	1. Simple error - "bad data"	1. Undetected calculation errors
2. Limits of precision	2. Imprecise orbital elements	2. Limits of precision	2. Imprecise cognitive parameters (e.g., priors and likelihoods)
3. Systematic bias in instruments	3. Imprecise planetary masses	3. Systematic bias due to task selection or psychometrics	3. Imprecise utility function
4. Inadequate corrections for known sources of systematic error	4. Insufficiently converged infinite-series calculations	4. Inadequate corrections for known sources of systematic error	4. Insufficiently approximated ideal solution/need for revision of ideal solution

5. Imprecise fundamental constants	5. Need for higher-order terms	5. Theory behind theory-mediated measurement incorrect	5. Need for different resource-rational algorithm
6. Not yet identified sources of systematic error	6. Forces not taken into account	6. Not yet identified sources of systematic error	6. Resource costs not taken into account
	7. Gravitation theory wrong		7. Universal optimality wrong